Resumen de Tesis Doctoral / *Resum de Tesi Doctoral*

Medidas Computacionales de Ganancia de Información y Refuerzo en Procesos de Inferencia

Mesures Computacionals de Guany d'Informació i Reforç en Processos d'Inferència.

***Título Original / Títol Original**:*
Computational Measures of Information Gain and Reinforcement in Inference Processes

José Hernández Orallo

Director: **Prof. Dr. Rafael Beneyto Torres**
Catedrático de Lógica y Filosofía de la Ciencia
Universitat de València

Septiembre 1999

Según el acuerdo tomado por la Comisión de Doctorado de fecha 21 de junio de 1995, la exposición, la defensa y la redacción de las tesis doctorales se podrá hacer en un idioma diferente de los oficiales de esta Universidad, siempre que la tesis tenga un amplio resumen (antecedentes, objetivos, discusión y conclusión) en uno de los dos idiomas oficiales.

Aunque el hecho de realizar este resumen sobre la tesis se debe a la obligación, desde mi punto de vista razonable, de escribir al menos un sumario en una de las dos lenguas oficiales de la Universitat de València, merece la pena aprovecharlo para realizar un resumen comprensible y comprensivo de la tesis, con el objeto de que pueda ser leído completamente e independientemente en una primera aproximación a la misma.

Es evidente que el hecho de resumir en una extensión de poco más de sesenta páginas una tesis de alrededor de trescientas implica dejar cosas fuera. No sólo se omitirán las demostraciones, gran parte del contenido introductorio, algunas aplicaciones y los apéndices, sino que también se dejan fuera ciertas reflexiones, ejemplos, cuestiones y detalles que sólo se pueden encontrar en el texto original (que en lo que sigue denominaré disertación). El resultado de este extracto, por tanto, tiene poco de carácter de tesis y más de artículo, en el que se hace hincapié principalmente en los objetivos, resultados y conclusiones.

Segons l'acord pres per la Comissió de Doctorat de data 21 de juny de 1995, l'exposició, la defensa i la redacció de les tesis doctorals es podrà fer en un idioma diferent dels oficials d'aquesta Universitat, sempre que la tesi tinga un ampli resum (antecedents, objectius, discussió i conclusió) en un dels dos idiomes oficials.

Encara que el fet de realitzar aquest resum sobre la tesi es deu a l'obligació, del meu parer raonable, d'escriure'n almenys un sumari en una de les dues llengües oficials de la Universitat de València, cal aprofitar-ho per a realitzar-ne un resum comprensible i comprensiu, amb l'objecte de que puga ser llegit completament i independentment en un primer apropament a la tesi.

És ben cert que el fet de resumir en una extensió de poc més de seixanta pàgines una tesi de vora tres-centes implica deixar-ne coses fora. No només s'ometran les demostracions, gran part del contingut introductori, algunes aplicacions i els apèndixs, sinó que també es deixen fora certes reflexions, exemples, qüestions i detalls que només es poden trobar al text original (a partir d'ara denominat dissertació). El resultat d'aquest extracte, per tant, té poc de caràcter de tesi i més d'article, al qual es recalquen principalment els objectius, resultats i conclusions.

Resumen y Palabras Clave

Esta tesis se centra en el estudio formal de la utilidad y resultados de la síntesis de conceptos inductivos y deductivos en términos de ganancia de información y refuerzo en sistemas de inferencia. El conjunto de medidas que se introducen permiten un análisis detallado y unificado del valor del resultado de cualquier proceso de inferencia con respecto a la entrada y el contexto (conocimiento previo o sistema axiomático).

Aunque las medidas más importantes, ganancia computacional de información, refuerzo e intensionalidad, se definen de manera independiente, permiten (solas o combinadas) formalizar y comprender mejor varias nociones que han sido tratadas tradicionalmente de una manera bastante ambigua: novedad, la diferencia entre explícito e implícito, informatividad, sorpresa, interés, plausibilidad, confirmación, comprensibilidad, 'consiliencia', utilidad e incuestionabilidad.

La mayoría de las medidas se aplican a diferentes tipos de teorías y sistemas, desde la estimación de la capacidad de predicción, la optimalidad de representación, o el poder axiomático de teorías lógicas, sistemas software y bases de datos, hasta la evaluación justificada de las habilidades intelectuales de agentes cognitivos y seres humanos.

Palabras Clave: *Procesos de Inferencia, Medidas de Evaluación, Inducción, Deducción, Información, Complejidad Kolmogorov, Razonamiento, Paradoja de la Inferencia, Ganancia de Información, Confirmación de la Inferencia, Refuerzo, Medición de Capacidades Cognitivas, Sistemas Basados en el Conocimiento, Aprendizaje Computacional, Programación Lógica Inductiva, Intensionalidad.*

Resum i Paraules Clau

Aquesta tesi se centra en l'estudi formal de la utilitat i resultats de la síntesi de conceptes inductius i deductius en termes de guany d'informació i reforç en sistemes d'inferència. El conjunt de mesures que s'introdueixen permeten una anàlisi detallada i unificada del valor del resultat de qualsevol procés d'inferència respecte a l'entrada i el context (coneixement previ o sistema axiomàtic).

Tot i que les mesures més importants, guany computacional d'informació, reforç i intensionalitat, es defineixen de manera independent, permeten (soles o combinades) de formalitzar o comprendre millor vàries nocions que han estat tractades tradicionalment d'una manera prou ambígua: novetat, la diferència entre explícit i implícit, informativitat, sorpresa, interès, plausibilitat, confirmació, comprensibilitat, 'consiliència', utilitat i inqüestionabilitat.

La majoria de les mesures s'apliquen a diferents tipus de teories i sistemes, des de l'estimació de la capacitat de predicció, l'optimalitat de representació o el poder axiomàtic de teories lògiques, sistemes software i bases de dades, fins l'avaluació justificada de les habilitats intel·lectuals d'agents cognitius i éssers humans.

Paraules Clau: *Processos d'Inferència, Mesures d'Avaluació, Inducció, Deducció, Informació, Complexitat Kolmogorov, Raonament, Paradoxa de la Inferència, Guany d'Informació, Confirmació de la Inferència, Reforç, Mesurament de Capacitats Cognitives, Sistemes Basats en el Coneixement, Aprenentatge Computacional, Programació Lògica Inductiva, Intensionalitat.*

Contenido / *Contingut*

1. Antecedentes

Razonar concierne diferentes procesos de inferencia. Desde Aristóteles, la inferencia deductiva ha ido asociada con la lógica, dedicada a discernir los razonamientos siempre válidos (o siempre inválidos) de aquellos contingentes. En cambio, el interés por la inferencia inductiva (o hipotética) es mucho más reciente (relativamente) y se ha dedicado principalmente a encontrar una metodología útil para la inducción y a establecer criterios de selección de hipótesis.

Durante la primera mitad del siglo XX, la lógica sufrió un proceso de formalización. En cambio, todos los intentos de 'logicizar' la inducción han fracasado (exceptuando soluciones parciales para la abducción) y sólo se han presentado formalizaciones de la inducción basadas en el cálculo de probabilidades o de la teoría de la información. Esta situación ha exagerado las diferencias entre deducción e inducción, hasta incluso hablarse de *dualidad* entre los procesos de inferencia. De este modo, sus objetivos y problemas han sido bien diferentes. La visión semántica (y mayoritaria) de la deducción, se dedica a examinar la completitud y corrección de diversos sistemas axiomáticos. Sin embargo, la cuestión prístina de la inducción ha sido siempre la de su justificación, por lo general mediante el establecimiento de criterios de selección de hipótesis, ya sean epistemológicos o metodológicos.

Los criterios que han sido más vindicados tanto en la filosofía de la inducción como en el área del aprendizaje automático son plausibilidad (o probabilidad), utilidad, simplicidad, comprensibilidad e informatividad, aunque muchos de ellos se pueden entender de maneras diferentes (e incluso contradictorias). Estos criterios se han adaptado a otros procesos de inferencia hipotéticos, como la abducción y la analogía, pero raramente se han considerado aspectos fundamentales para la deducción. La razón, a primera vista, es simple: un demostrador de teoremas *no* debe catalogar sus teoremas, se debe dedicar *exclusivamente* a decir qué fórmulas son teoremas y cuáles no.

Sin embargo, la deducción, como proceso de inferencia es mucho más que establecer qué es lo correcto y qué es posible deducir a partir de unos axiomas y una reglas de derivación. Cualquier proceso de inferencia se define como un "proceso de razonamiento por el cual un agente modifica parte de sus creencias". Existen, por tanto, muchos puntos en común entre todos los procesos de inferencia. En primer lugar, existe un rasgo *intencional*, con el objetivo de obtener nuevo conocimiento a partir de nuevos datos y/o creencias previas, ya sean reales o imaginarias. En segundo lugar, existe un carácter *epistemológico* de cada inferencia, porque el resultado es *nueva* información, que no era conocida de manera explícita antes del proceso de inferencia. Este resultado puede ser un concepto (un hecho, una regla o propiedad), o la refutación o confirmación de una creencia previa o asumida. Cualquier

asignación intermedia para una creencia, entre la refutación y la confirmación, es asimismo posible, y se produce por el grado de refuerzo que diferentes procesos de inferencia le han ido asignando a partir de otras creencias y sus respectivos grados de credibilidad o plausibilidad. Es de resaltar que, en este contexto, las características fundamentales para entender los procesos de inferencia vienen a ser las nociones de información, novedad, creencia, explicitez y refuerzo.

1.1 Información y la Paradoja de la Inferencia

En principio, parece claro que la inducción debe, de algún modo, depender de la deducción, porque cualquier hipótesis debe comprobarse deductivamente con la evidencia. Sin embargo, durante la última mitad del siglo veinte se ha intentado presentar a la inducción como la inversa de la deducción, en términos de ganancia de información. En otras palabras, la visión de que la inducción aporta nuevo conocimiento o *incrementa información* mientras que la deducción *la decrementa* se ha hecho bastante popular. A ella ha contribuido en gran medida la interpretación lógica de la probabilidad (o la interpretación probabilística de la lógica), iniciada por John Maynard Keynes [Keynes 1921] y Rudolf Carnap [Carnap 1950, 1952], aunque algunas ideas básicas se pueden encontrar ya en los trabajos de George Boole. Esta visión fue axiomatizada por Bar-Hillel y Carnap [Bar-Hillel and Carnap 1953], que definieron la información de una sentencia lógica como el logaritmo negativo de su probabilidad, dada por la interpretación probabilística del cálculo de predicados de primer orden.

Pero esta relación hacía muy difícil el armonizar las teorías semánticas de la información con la teoría de la información estadística. En concreto, la asignación habitual $I(P) = -\log p(P)$, originaria de la teoría de Shannon, es hoy en día comúnmente utilizada en diferentes campos del aprendizaje automático, especialmente en la Programación Lógica Inductiva (ILP) (véase e.g. [Muggleton 1996]). A partir de aquí, si $P \models Q$, entonces, bajo el Cálculo Probabilístico de Carnap, se tiene que $p(P) \leq p(Q)$. Utilizando la asignación anterior, se obtiene que $I(P) \geq I(Q)$, es decir, la premisa es más informativa que la consecuencia.

A partir de este razonamiento, la afirmación popular "la deducción pierde información mientras que la inducción la incrementa" parece convincente. Pero más aún, la teoría de la información de Shannon ha sido substituida por la moderna visión descriptiva de la información, que subsume a la visión estadística de la información. La información descripcional, complejidad algorítmica o complejidad Kolmogorov ha permitido re-entender la inducción como compresión de información. De este modo se ha hecho muy popular la visión moderna de la navaja de Occam basada en la complejidad de Kolmogorov, denominada el principio de la longitud de la descripción mínima (Minimum Description Length, MDL). El principio MDL, elige la teoría que minimiza $I(T \mid E)$ siendo T la hipótesis y E la evidencia, para maximizar

la probabilidad $p(T \mid E)$. Ya que la evidencia viene dada y se asume correcta, $I(E) = 0$, obtenemos, mediante aplicación de logaritmos a la regla de Bayes, la siguiente expresión:

$$I(T \mid E) = I(T) + I(E \mid T)$$

Hasta ahora esta expresión es sólo la fórmula probabilística (o informacional) tradicional de la inferencia inductiva[1]. El paso *particular* del principio MDL es aproximar la función $I(x|y)$ por $K(x|y)$, la complejidad Kolmogorov relativa de x respecto de y [Muggleton et al. 1992] [Muggleton and Page 1994]. Como, por lo general, la complejidad descripcional $K(E \mid T)$ es baja (cómo describir la evidencia a partir de la teoría), el resultado es que minimizar $I(T \mid E)$ supone minimizar $K(T)$. Es decir, cuanto más corta sea la hipótesis, mejor. En realidad $K(E \mid T)$ es sólo despreciable si tenemos toda la evidencia que produce una teoría, porque en el caso de que dicha evidencia sea un subconjunto debemos describirlo de alguna manera. La visión general del principio MDL [Rissanen 1978, 1986, 1996] se define, por tanto, de la siguiente manera:

La mejor teoría para explicar un conjunto de datos es la que minimiza la suma de:

- *la longitud, en bits, de la descripción de la teoría; y*
- *la longitud, en bits, de los datos cuando se codifican a partir de la teoría.*

En segundo lugar, se añaden las excepciones, si las hubiera.

Aunque el principio MDL ha dado muy buenos resultados en la práctica, se obtienen resultados poco intuitivos al usar $K(x)$ en vez de $I(x)$. Por ejemplo, si consideramos dos hipótesis P y Q, siendo P más simple que Q (es decir, $K(P) < K(Q)$), y además se tiene que $P \models Q$, entonces, a partir de la correspondencia entre información y probabilidad tendríamos $p(P) > p(Q)$ pero, en cambio, a partir del Cálculo Probabilístico de Carnap, se tiene que $p(P) \leq p(Q)$. El resultado es que la combinación de la deducción y la inducción no es posible, porque los criterios de plausibilidad de la inducción divergen claramente con los criterios de plausibilidad (o probabilidad) de la deducción.

Aparte de existir demasiadas suposiciones y simplificaciones para llegar a ciertos principios válidos en la práctica, existe una cuestión de fondo que impide que deducción e inducción puedan compartir una medida de información unificada para los dos. Esta cuestión de fondo es casi tan vieja como la lógica y se conoce como la "paradoja de la inferencia", expresada de esta manera tan rotunda por [Cohen and Nagen 1935]: "*si la conclusión de una inferencia no está contenida en las premisas, no puede ser válida, y si no es diferente de ellas, es inútil; sin embargo, la conclusión no puede estar contenida en las premisas y ser al mismo tiempo novedosa; consecuentemente, las inferencias no pueden ser al mismo tiempo válidas y útiles*". Esta paradoja es todavía "el mayor problema sin resolver

[1] Paso que ya es discutible, puesto que, bajo la correspondencia $I(P) = -\log p(P)$, el hecho de maximizar $p(T \mid E)$ obliga a minimizar $I(T \mid E)$, que no es más que forzar a que la teoría sea lo menos informativa posible respecto a la evidencia. Este problema se discutirá más adelante y es uno de los que se intentan abordar por la presente tesis.

para la justificación de la deducción" [Dummett 1973] o, en palabras de Hintikka, "el escándalo de la deducción" [Hintikka 1973], pues niega a la deducción la categoría de proceso útil y valioso.

Es importante resaltar que la fuente tradicional acerca de esta paradoja es Mill, que le motivó a buscar en la inducción el proceso de inferencia "informativo" que faltaba. La raíz de la cuestión reside en reconocer que esta paradoja de la deducción afecta (y de hecho está estrechamente relacionada) con la paradoja de la inducción (o imposibilidad de la probabilidad inductiva). En otras palabras, el escándalo de la deducción y de la inducción reside en no reconocer que la paradoja de la inducción y de la deducción son dos caras de la misma moneda.

Veamos cómo se relacionan ambas paradojas. Para ello recurriremos al argumento de Popper y Miller. Popper y Miller comenzaron en 1983 un debate vigoroso (véase [Mura 1990] para un seguimiento exhaustivo) sobre la conexión entre las relaciones deductivas y el soporte probabilístico en su papel "*Proof of the Impossibility of Inductive Probability*" [Popper and Miller 1983]. Popper y Miller sostienen que cualquier soporte probabilístico positivo de una evidencia e para una hipótesis h, medido por $s(h,e) = p(h,e) - p(h)$ se debe exclusivamente a relaciones deductivas (entendidas apropiadamente) entre e y h. Un corolario inmediato de esto es que el soporte inductivo (es decir, no deductivo) no existe. En otras palabras, Popper y Miller mantienen que todo el soporte probabilístico es deductivo.

Su resultado se basa en que "lo que queda de h una vez que le hemos quitado todo lo que está implicado por e, es una proposición que en general es contra-dependiente respecto a e" [Popper and Miller 1983]. Más claramente, es fácil demostrar por lógica elemental (véase el capítulo 1 de la tesis o [Cussens 1998]) que:

Corolario 1.1 (*en la disertación 1.3*) b es deductivamente independiente de a si y sólo si $\neg a \models b$.

Por tanto, el soporte inductivo puro parece ser imposible. Para que la dependencia entre a y b sea puramente inductiva, necesariamente tiene que darse el caso de que a y b sean deductivamente independientes, lo que es equivalente a tener $\neg a \models b$.[2] Esto rebate la visión moderna de inducción y deducción como procesos inversos: "*cualquier noción de inducción como una clase de complemento a la deducción resulta insostenible*" [Cussens 1998]. Sin embargo, "*no hay razón para suponer que la inferencia inductiva no puede estar contaminada deductivamente. Puede haber una relación entre deducción e inducción, sin los dos tipos de inferencia ser equivalentes, o uno reducible al otro.*" [Cussens 1998].

A pesar de la rotundidad del razonamiento, no ha dejado de sorprender a aquéllos que creían firmemente en la inducción como proceso ampliativo y en la deducción como proceso recuperativo de la información. No obstante, parte de esta sorpresa (o incluso incredulidad) se puede deber a que se pasa por alto que todo el razonamiento anterior está asumiendo implícitamente la omnisciencia de la lógica. En otras

[2] Es de remarcar que, desde un punto de vista descripcional, es decir $K(x)$, la dependencia y la independencia deductivas, concretamente $a \models b$ y $\neg a \models b$, están extremadamente cerca.

palabras, asume que deducción e inducción se realizan en un sistema lógico completo, por el cual se puede obtener siempre si dos enunciados son deductivamente dependientes o independientes.

Además, se asume la versión más fuerte de la omnisciencia incluso para lenguajes universales: no sólo se ignora la posibilidad de error en la deducción, el tiempo y el esfuerzo necesarios para establecer cualquier relación deductiva sino que en la interpretación de los resultados se supone que dicha relación siempre se va a poder establecer, ignorando los resultados de incompletitud de Gödel.

Pero incluso en un sistema restringido completo, la visión de la deducción como omnisciente es incompatible con la visión original de ganancia de información. Cualquiera que haya trabajado en deducción desde un punto de vista pragmático, es decir un punto de vista no omnisciente, sabe perfectamente que muchas inferencias deductivas aportan *información* dentro del mismo sistema axiomático, sin cambiar la semántica (el conjunto de hechos inferibles). Más aún, merece la pena recordar y mantener explícita esta información, porque se ha invertido un cierto esfuerzo en los resultados. La mayoría de las nuevas lógicas aparecidas en la segunda mitad del siglo XX, incluso las no-monótonas, no han considerado esta cuestión importante, la cual devolvería la deducción al marco del resto de procesos de inferencia.

Este mismo problema se ha presentado incluso en el campo de la lógica modal y epistémica, y parte del fracaso de muchas de estas lógicas se debe a suponer la omnisciencia. Sin embargo, si eliminamos los axiomas de omnisciencia, los sistemas lógicos que resultan son muy pobres y prácticamente inservibles. La solución, como propone [Duc 1997], se debe basar en considerar la lógica de una manera dinámica, potencial, …. "*Las leyes de la lógica son lo que el agente sabe implícitamente; él no tiene por qué poseerlas permanentemente*" [Duc 1997]. Para ello es necesario reconocer que es esta dinámica del razonamiento la que va haciendo explícito lo que las leyes de la lógica marcan implícitamente. Por tanto, el agente *puede* conocer un cierto resultado que está implícitamente en sus creencias anteriores si *piensa lo suficiente*. Tras este esfuerzo, el agente obtiene nueva información que no conocía explícitamente.

Esta idea de la deducción como un proceso informativo fue reivindicada fundamentalmente por Hintikka (véase por ejemplo [Hintikka 1970a]). Dentro de la lógica de primer orden, introdujo la diferencia entre información profunda e información superficial, distinguiendo claramente entre lo que se sabía explícitamente en un momento dado (información superficial) y lo que se sabía implícitamente o era posible inferir (información profunda). Con su teoría justificó la introducción de conceptos auxiliares en un sistema axiomático y la necesidad de reorganización de los mismos para hacer los sistemas deductivos más manejables y útiles.

La teoría de la información semántica de Hintikka, aunque filosóficamente muy influyente, tenía bastantes problemas prácticos. Se limitaba a la lógica de primer orden, dependía del número de individuos del universo de discurso y nunca llegó a extenderse para incluir a la inducción de una manera uniforme.

Además, la teoría de Hintikka se desarrolló al mismo tiempo que empezaban a popularizarse en Europa Occidental y América (por medio de Chaitin) las nociones de información algorítmica y complejidad descripcional introducidas por Kolmogorov y Solomonoff.

Aunque existe una abundante literatura sobre inducción y complejidad de Kolmogorov, hasta la fecha, y por lo que conoce el autor de esta tesis, no existe una aproximación basada en la complejidad descripcional para el problema de la paradoja de la deducción o, más concretamente, del estudio de cuán informativa (descripcionalmente hablando) es una conclusión a partir de unas premisas.

Una de las ideas fundamentales de esta tesis es por tanto considerar la dependencia entre premisas y conclusiones de una manera descriptiva y no semántica. Sin embargo, el uso de la teoría de complejidad *absoluta* de Kolmogorov, denotada por la función *K*, no lleva tan fácilmente a resultados intuitivos (y quizás ésta sea una de las razones por la que nunca se haya aplicado al problema de la deducción). Más concretamente, si se considera $a \vDash b$ entonces $K(b \mid a)$ suele ser muy bajo (o incluso 0 si *b* son todas las consecuencias lógicas de *a*) mientras que $K(a \mid b)$ suele ser un valor alto. En otras palabras, si el tiempo de computación no se considera (como en la deducción omnisciente), el resultado clásico se obtiene una vez más: la deducción no suele aportar información mientras que la inducción sí.

Las cosas cambian radicalmente si introducimos el tiempo y consideramos alguna variante espacio-temporal de *K*. Como veremos, si se considera la variante de Levin, denotada por la función *Kt*, vemos que tanto $Kt(b \mid a)$ como $Kt(a \mid b)$ pueden ser mayores que 0, puesto que si el sistema deductivo tiene recursos limitados (espacio o tiempo), la inferencia deductiva de *b* a partir de *a* proporciona información, dependiendo de cuán explícita sea la relación $a \vDash b$ en el sistema.

1.2 Recursos y Esfuerzo de Inferencia

Los resultados de Gödel y las nociones de (in)computabilidad de Church y Turing representan los resultados más importantes acerca de lo que se puede deducir y lo que se puede computar. Sólo es a partir del advenimiento de las primeras computadoras que comienza a nacer un interés por saber qué es deducible y computable *de una manera eficiente*, algo que había sido secundario hasta bien entrado el siglo XX.

Los resultados han sido en la mayor parte desalentadores. La deducción (más concretamente el problema de la satisfacibilidad) es intratable (de complejidad no polinómica) incluso para la lógica proposicional. El problema de la intratabilidad ha acompañado a los sistemas de deducción automática (o demostradores automáticos) desde sus inicios. Del mismo modo, los prototipos y sistemas de inducción automática también se han encontrado (incluso en mayor grado) con el problema de la intratabilidad.

No ha sido hasta el último cuarto del siglo XX que un nuevo paradigma de inferencia o razonamiento está tomando fuerza, denominado razonamiento limitado por recursos (*resource-bounded reasoning*). Este paradigma fue introducido por I.J. Good [Good 1971] y H. Simon [Simon 1982] como una manera de hacer factible el razonamiento en aplicaciones de inteligencia artificial de media y gran escala [Russell and Wefald 1991]. La corriente más técnica de esta propuesta evolucionó a partir de la noción teórica de algoritmo aproximado (o *anytime*). "Los algoritmos *anytime* ofrecen un compromiso entre tiempo de computación y la calidad de los resultados" [Zilberstein 1995]. La razón de este compromiso es obvia; para algunos problemas "*no es factible (computacionalmente) o deseable (económicamente) computar la respuesta óptima*" [Zilberstein 1996]. En el caso de la deducción, bajo este contexto, la idea no es encontrar la demostración de un teorema sino obtener su grado de plausibilidad. Al inspirarse en técnicas aleatorias, como los métodos de Montecarlo o algoritmos genéticos, el resultado es una respuesta probabilísticamente correcta, cuya probabilidad depende del tiempo que se ha suministrado al algoritmo. Por ejemplo, existe un algoritmo aproximado que da con una probabilidad increíblemente alta de éxito si un número es primo, y se pueden usar en la práctica en lugar de otros algoritmos exactos pero computacionalmente más costosos.

No obstante, es necesario definir una nueva medida de optimalidad. En general, la "*Optimalidad (...) se define con respecto al conocimiento del sistema y sus capacidades computacionales*" [Zilberstein 1999]. Para una medida de optimalidad es necesario evaluar el esfuerzo computacional de lo que se sabía antes de un proceso de inferencia y lo que se sabe después de él. En otras palabras, se requiere una medida de *ganancia de información* dependiente de los recursos. Desde un punto de vista computacional, los recursos más interesantes a considerar son la memoria y el tiempo.

Aparte del compromiso de utilización de los recursos y la calidad de la solución, existe otra importante cuestión a la hora de gestionar los recursos computacionales: cuándo se invierten los recursos, es decir la distribución temporal de los mismos. Algunos procesos de inferencia, como la analogía o la abducción, trabajan *al vuelo*, es decir, son *perezosos* (*lazy*) [Aha 1997], en el sentido de que se usan sólo cuando se necesitan, y otros procesos son más *anticipativos* (*eager*), en el sentido de que intentan obtener conceptos o reglas que puedan ser necesarios en el futuro, como los que genera la inducción constructiva.

Más concretamente, en el caso de la inducción, los métodos de aprendizaje anticipativos (*eager*) extraen toda la regularidad de los datos para trabajar con conocimiento intensional, es decir, un modelo. Ejemplos de aprendizaje anticipativo son el razonamiento basado en modelo (*Model Based Reasoning*, MBR) y la programación lógica inductiva (Inductive Logic Programming, ILP). El principio MDL se usa muy a menudo en estos dos campos aunque muchas veces se ignora que este principio da una teoría que usualmente es anticipativa para datos comprimibles pero perezosa para datos incomprimibles. En esta tesis se investigarán criterios más

anticipativos (o siempre anticipativos), que inviertan en teorías más complejas (o más intensionales).

La distinción entre anticipativo y perezoso, sin embargo, no ha sido establecida de una manera clara en la literatura para el caso de la deducción. No obstante, la deducción es también a veces perezosa, como las inferencias deductivas cotidianas, y algunas veces anticipativa, como la práctica matemática. Solamente las técnicas de especialización y transformación de programas [Pettorossi and Proietti 1990, 1996a, 1996b] [Dershowitz and Reddy 1992] se ocupan de la transformación de sistemas deductivos o programas en otros más eficientes, preparando la representación del programa para los hechos esperados que debe cubrir, algo que puede verse como deducción anticipativa.

El tiempo de reacción de la inferencia inductiva y deductiva es crucial en sistemas activos, y el cociente entre tiempo de reacción y calidad de respuesta es el punto más importante, ya resaltado por [Horvitz 1990]. Aquí se conjugan las técnicas de razonamiento de recursos limitados y una apropiada elección entre técnicas perezosas y anticipativas, para tener la representación del conocimiento de la manera óptima con el fin de acelerar el tiempo de respuesta del tipo de problemas o casos que se esperan. Atendiendo a esta expectación (en gran medida predicha a partir de los casos pasados), se podría definir una medida de optimalidad de representación.

La elección entre inferencia perezosa y anticipativa depende claramente de los recursos temporales y espaciales que se hayan consumido (el esfuerzo o ganancia de información computacional) pero también de la frecuencia de uso y el grado de plausibilidad de ciertas partes del conocimiento. Este segundo factor será aproximado por la teoría del refuerzo, cuyos precedentes veremos en la próxima sección. Ambos aspectos determinan que alguna información merece mantenerse de manera explícita (información intermedia) mientras que otra puede (y debe) olvidarse. En otras palabras, se requiere un criterio de olvido para discernir qué conocimiento debe mantenerse explícito.

La diferencia entre explícito e implícito está relacionada con la diferencia entre definiciones por extensión y por intensión (o por comprensión). Qué conceptos recordar y de qué manera (extensionalmente o intensionalmente) determinan también el esfuerzo al recordarlos o usarlos. Asimismo, muchos aspectos de la inteligibilidad y comprensibilidad de conceptos están también relacionados [Sommer 1995], aunque sólo se han introducido de manera informal. Una formalización de estos aspectos permitiría medir la complejidad de (o dificultad de comprender) un concepto, y podría servir para evaluar sistemas (artificiales o no).

1.3 Inferencia y Confirmación

Otro aspecto importante de cualquier proceso de inferencia es que una inferencia puede confirmarse o refutarse. Incluso en el caso de inferencia no-hipotética (la

deducción clásica), es completamente diferente afirmar que "*B* es una consecuencia lógica de *A*" que aseverar que "*B* es una consecuencia lógica de *A* debido a la demostración *D*". Esto muestra que un demostrador de teoremas realmente proporciona información útil, porque una demostración aporta nuevo conocimiento, meta-conocimiento acerca de la veracidad de otras partes de conocimiento. De hecho, las matemáticas están llenas de conjeturas, que podrían o no podrían ser refutadas. Pero incluso en el caso de la deducción computacional se debe admitir alguna posibilidad de error, y, por consiguiente, las confirmaciones adicionales son útiles. Evidentemente, para la inferencia hipotética, el papel de la confirmación es más evidente porque los criterios de evaluación no son suficientes para seleccionar la hipótesis 'correcta' con certeza. Sin embargo, cualquier confirmación procede de conocimiento que se origina directamente o indirectamente de la evidencia, con lo que el problema del criterio de evaluación es inevitable.

Recapitulemos dos soluciones diferentes para el problema de la confirmación. Dos filósofos y lógicos del Wiener Kreis afrontaron el problema: Carnap y Hempel. Carnap desarrolló un concepto cuantitativo del grado de confirmación para una hipótesis dada una evidencia, como un valor entre 0 y 1, y lo asoció con una noción de probabilidad. Por el contrario, Hempel introdujo un concepto cualitativo de confirmación, es decir, una relación booleana entre la hipótesis y la evidencia, en el sentido que *E* confirma *H* o *E* no confirma *H*. Con el fin de pergeñar una relación lógica de confirmación, Hempel introdujo cinco condiciones de idoneidad (*adequacy*) [Hempel 1943, 1945]. Son las siguientes (de [Flach 1995]):

(H1) *Condición de Implicación*[3]: cualquier sentencia que es implicada por una observación es confirmada por ella.

(H1.1) Cualquier observación se confirma a sí misma.

(H2) *Condición de Consecuencia*: si una observación confirma cada una de las sentencias de una clase *K*, entonces confirma cualquier sentencia que es una consecuencia de *K*.

(H2.1) *Condición especial de consecuencia*: si una observación confirma una hipótesis *H*, entonces también confirma cualquier consecuencia de *H*.

(H2.2) *Condición de equivalencia*: si una observación confirma una hipótesis *H*, entonces confirma también toda hipótesis que es lógicamente equivalente con *H*.

(H2.3) *Condición de conjunción*: si una observación confirma dos hipótesis cualesquiera, entonces confirma su conjunción.

(H3) *Condición de consistencia*: toda observación consistente lógicamente es compatible lógicamente con la clase de todas las hipótesis que confirma.

(H3.1) A no ser que una observación sea auto-contradictoria, no confirma ninguna hipótesis con la cual no sea lógicamente compatible.

(H3.2) A no ser que una observación sea auto-contradictoria, no confirma cualesquiera hipótesis que se contradigan entre sí.

[3] He traducido 'entailment' por implicación para facilitar la lectura de las condiciones de idoneidad, aunque más bien se debería traducir por necesidad lógica.

(H4) *Condición de equivalencia para las observaciones*: si una observación *B* confirma una hipótesis *H*, entonces cualquier observación equivalente con *B* también confirma *H*.

(H5) *Condición de consecuencia inversa*: si una observación confirma una hipótesis *H*, entonces también confirma cualquier fórmula que implique *H*.

Informalmente, H1 y H2 pueden identificarse como confirmaciones deductivas (hacia abajo), H3 es una confirmación inductiva en el sentido de Popper (la teoría no ha sido todavía refutada por la evidencia), y H5 es una confirmación abductiva (hacia arriba). H4 es la más natural e indudable, al menos si no se tienen en cuenta modalidades. Sin embargo, H2 y H2.1 resultan ser inconsistentes con H5, un problema denominado "la paradoja de la confirmación" [Hempel 1943, 1945] [Hesse 1974]. La solución de Hempel es eliminar una de las dos condiciones, pero, como afirma Flach, "*Hempel resuelve el problema en el nivel formal al eliminar la condición de consecuencia inversa en favor de la condición de consecuencia. No obstante, en el nivel intuitivo la paradoja se mantiene, puesto que Hempel no proporciona una justificación clara acerca de su elección*" [Flach 1995]. Más aún, H2.2 genera algunos problemas con fórmulas generales que Hempel intenta resolver a través de una relación más estrecha de confirmación directa, que está, de alguna manera, relacionada con el principio del subconjunto, es decir, si dos hipótesis cubren los datos, elige la más específica. La solución cualitativa de Flach es mucho más convincente; él separa dos subconjuntos de condiciones de idoneidad que dan cuenta separadamente de los razonamientos, según él, explicativos (abductivos) y confirmativos (descriptivos), de este modo subrayando la distinción entre razonamiento abductivo e inductivo. Sin embargo, en mi opinión, la visión original de la confirmación no está representada por la segunda elección únicamente, ni tampoco por la primera. El problema es que un modelo cualitativo de la confirmación no puede conciliar H2 con H5, es decir, la confirmación hacia abajo o hacia adelante (deductiva) con la confirmación hacia arriba o hacia atrás (abductiva), porque ambas tienen diferente fuerza.

Un argumento de esta tesis, que es especialmente respaldado por los resultados del capítulo 5, es que es posible ponderar consistentemente ambas fuentes de confirmación. Esto no podría hacerse con una medida de probabilidad, en un sentido estricto, sino que debe hacerse con una medida de plausibilidad, que se desmarca claramente del cálculo probabilístico de Carnap. Otra razón adicional es evitar el problema de la no-informatividad de las aproximaciones probabilísticas de la confirmación, como ya apuntara Popper: "*Aquellos que identifican la confirmación con la probabilidad deben creer que un alto grado de probabilidad es deseable. Ellos aceptan implícitamente la regla: 'Siempre elige la hipótesis más probable!' Ahora puede demostrarse que esta regla es equivalente a la siguiente regla: '¡Siempre elige la hipótesis que va lo menos allá posible de la evidencia como sea posible!*" ([Popper 1963], p.p. 289-90), o quizás, coge la evidencia misma como una hipótesis completamente extensional. Carnap, por el contrario, ignora este problema mediante la separación del problema de la probabilidad con el interés: "*La lógica inductiva sola no determina y no puede determinar la mejor hipótesis sobre una*

evidencia dada… Esta preferencia se determina por factores de muchos tipos diferentes..." ([Carnap 1950], p.221, a partir de [Flach 1995], p. 30). Es posible que un *grado de intensionalidad* o *una medida de ganancia de información* pudieran estar en la mente de Carnap.

De un modo más preciso, la aproximación de una teoría de confirmación que se acometerá en el capítulo 5 está basada en una propagación gradual (no booleana) de la confirmación, una solución entre la de Hempel y la de Carnap, que permite incluir tanto H2 como H5, una teoría que está entre el principio MDL y el criterio de informatividad de Popper. Se mostrará cómo esta medida de refuerzo es útil para la deducción, inducción y abducción.

1.4 El Problema de la Combinación de Procesos de Inferencia

La investigación en inteligencia artificial ha estudiado los procesos de inferencia de una manera separada. Aunque la abducción ha sido vista algunas veces en conjunción con la deducción en modelos no monótonos de razonamiento o lógicas probabilísticas, la inducción, como la manera de generar teorías a partir de hechos o aprender de una manera automática, ha sido generalmente una cosa aparte, emprendida por la comunidad de aprendizaje automático.

Aparte de la visión de Popper y Miller de que todo soporte inductivo es deductivo, ha habido otros intentos de ver la inducción como deducción, con la ilusión de que todos los problemas de combinación se resolverían, porque habría un único proceso de inferencia. [Shanahan 1989] estudió el uso de la deducción para la predicción y la abducción para la explicación, y [Gregoire and Saïs 1996] sostienen que el razonamiento inductivo es a veces deductivo. En mi opinión, estos resultados se obtienen por conceptos erróneos o concepciones diferentes (típicas en IA) de alguno de los métodos de inferencia implicados, que, en cualquier caso, no resuelven el problema principal de su combinación. También ha influido mucho en estas visiones la confusión entre deducción y computación (también habituales en IA), ya que cualquier proceso de inducción que se quiera hacer sobre una máquina tiene que expresarse necesariamente en términos computacionales (que no deductivos).

En la última década, los primeros éxitos del aprendizaje computacional han motivado el uso de algunas de sus técnicas para otros problemas con un carácter más deductivo, como son la ingeniería del software, los sistemas de información (bases de datos) y la deducción automática [Langley and Simons 1995]. Pero sólo recientemente, la teoría de agentes ha abordado el problema del razonamiento combinando procesos de inferencia, al menos de una manera informal. Existe de nuevo cierto interés sobre la asociación de diferentes procesos de inferencia con el fin de hacer sistemas más inteligentes. De hecho, se ha admitido que muchos procesos de aprendizaje o de adquisición de conocimiento se pueden explicar como

combinaciones apropiadas de procesos de inferencia básicos: inducción, abducción y deducción.

En [Michalski 1993], diferentes combinaciones y variantes de razonamiento hipotético llevan a diferentes tipos de combinaciones de inferencia (véase la tabla 2.1 en el capítulo 2): generalización inductiva empírica, generalización inductiva constructiva (generalización + derivación deductiva), especialización inductiva, concreción, generalización abductiva constructiva (generalización + abducción). Esto da lugar a que los procesos de inferencia clásicos se pueden 'camuflar' bajo diferentes nombres.

La cuestión es la siguiente: si estos procesos de inferencia complejos o derivados se componen de procesos de inferencia básicos o más simples, ¿significa esto que la plausibilidad, informatividad y confirmación deben asignarse como la composición de sus partes? Y, si éste es el caso, ¿es posible combinar los criterios de plausibilidad de la inducción con los criterios de plausibilidad de la deducción no monótona? ¿Tiene el mismo significado la informatividad para la abducción, la inducción y la deducción? ¿Tiene algo que ver la complejidad (o comprensibilidad) de un problema deductivo con uno inductivo? Las cosas parecen más complicadas si pretendemos medir el valor, la novedad o la utilidad interna de estas inferencias, dependiendo del conocimiento anterior, el cual también ha sido construido a partir de procesos de inferencia variados.

El primer campo que requiere este tipo de medidas es el campo de la demostración automática de teoremas (ATP), que está muy interesada en evitar inferencias inútiles y mantener explícitas aquellas propiedades que pueden ser útiles para encontrar la demostración de un teorema.

El segundo campo es el de los agentes racionales de recursos limitados, donde el coste de la inferencia deductiva es un factor que debe ser minimizado por muy diversos medios, como evitar la rederivación de hechos útiles y comunes que son costosos de derivar a partir de los axiomas. Un sistema racional visto de una manera dinámica [Girard et al. 1989], debe distinguir qué es potencialmente derivable, qué es derivable de una manera factible, y qué es conocido explícitamente en una situación dada.

El tercer campo (o campos) es el de los sistemas de conocimiento, que engloba los sistemas software avanzados o adaptativos, los sistemas expertos de segunda generación y las bases de datos con descubrimiento automático de conocimiento, que cada día comienzan a integrar técnicas inductivas con el fin de adaptarse a nuevas situaciones o a los cambios del entorno sin modificaciones por parte del programador o del usuario.

Finalmente, existe una herencia en Filosofía y sobre todo en IA de estudiar la deducción y la inducción de una manera separada. Incluso muchos trabajos son insostenibles si se contrastan con la contrapartida deductiva/inductiva en términos

de información. Parece de este modo atractivo el estudiar una posible 'conciliación' entre inducción, deducción, confirmación, uso e información.

Ha habido, por supuesto, algunos intentos para resolver este problema. Los más exitosos, aunque en dominios restringidos, se basan en combinaciones de aprendizaje por refuerzo y optimizaciones de esfuerzo o recursos [Schmidhuber et al. 1997a, 1997b] [Martin 1998]. Sin embargo, el primer sistema que aborda el problema desde un punto de vista simbólico es SOAR [Newell 1990], pero se basa exclusivamente en un método de aprendizaje llamado *chunking*. La característica más importante de SOAR es que resalta cuándo y por qué se debe disparar un proceso de razonamiento. El sistema THEO [Mitchell et al. 1991] es también un sistema que se mejora a sí mismo y ya integra más métodos de aprendizaje. El sistema Noos [Arcos and Plaza 1996] combina técnicas de resolución de problemas (deductivas) con múltiples métodos de aprendizaje [Plaza and Arcos 1993] (inducción constructiva y razonamiento basado en casos [Armengol and Plaza 1994]).

Sin embargo, no se han presentado teorías que puedan dar cuenta de esta combinación en general. Para que estas combinaciones se puedan extender genéricamente a lenguajes de representación más potentes hacen falta criterios de evaluación que sean igualmente válidos y consistentes con todos los procesos de inferencia involucrados.

La mayoría de criterios existentes hasta la fecha se han presentado para un único proceso de inferencia. Aunque la mayoría no son aprovechables (o incluso inconsistentes) con el resto de procesos de inferencia, veamos brevemente cuáles son los más importantes.

Desde el punto de vista de la plausibilidad, muchos criterios de selección han sido defendidos. El criterio de simplicidad, como hemos visto, ejemplificado por la navaja de Occam y su moderna formalización bajo el principio MDL, es el más popular, fundamentalmente porque es al mismo tiempo un criterio de plausibilidad y metodológico, ya que la mayoría de resultados de tratabilidad se obtienen limitando el tamaño de las hipótesis. Otros criterios relacionados son la validación cruzada (cross-validation), estimadores de máxima probabilidad (Maximum Likelihood Estimators) atendiendo a alguna distribución asumida, el criterio de generalidad, de especificidad, de poder explicativo, etc.

Otros criterios como la dificultad, la necesidad o naturaleza de un concepto, la comprensibilidad han sido estudiados de una manera informal, o limitados a lógica de primer orden.

Para el caso de la deducción, la escena es sensiblemente peor. La deducción ha carecido de interés en ver qué deducciones eran mejores que otras, simplemente se trataba de ver cuáles eran correctas y cuáles no. Aparte de la aproximación de Hintikka, sólo existe un interés renovado en la evaluación de la inferencia deductiva desde precisamente los mismos campos que hemos comentado anteriormente, demostración automática de teoremas, agentes de racionalidad limitada y sistemas de

conocimiento. Se necesitan, por tanto, medidas para evaluar los conceptos auxiliares, la información intermedia, los lemas, etc., que tan útiles son en los problemas de deducción *genuina* y, sobre todo, imprescindible a media o gran escala.

Una pregunta instantánea antes de marcarse unos objetivos es qué razones había para suponer que en este momento las posibilidades de conseguir estas medidas unificadas eran más factibles que hace unos años. Desde el tiempo en que Hintikka ya intentó clarificar las paradojas (o escándalos) de la inducción y la deducción en términos de ganancia de información y utilidad, la visión generalmente aceptada de información ha cambiado de una manera muy importante, hacia la visión más general y universal basada en la complejidad de Kolmogorov. Además, algunos campos de la IA, como el aprendizaje automático o los sistemas basados en conocimiento han ayudado a clarificar el problema de la combinación de deducción e inducción y han dado relevancia al esfuerzo computacional y al coste de recursos de cualquier sistema con capacidades de raciocinio. Todo esto permite aspirar a unas metas más desafiantes con una posibilidad razonable de éxito.

Como conclusión de los antecedentes y motivaciones de esta tesis, no se debe perder de vista algo obvio y no por eso siempre presente en muchos trabajos acerca de razonamiento automático: razonar es mucho más que la demostración de unos teoremas, mucho más que la generalización inductiva y mucho más que la abducción, la analogía y otros procesos parciales de inferencia. Y también es mucho más que la suma de todos ellos. Si somos capaces de combinar consistentemente y provechosamente los diferentes procesos de inferencia, engrandeceremos el poder y las aplicaciones de los avances que, por separado, han tenido lugar en diferentes campos de la lógica, la filosofía de la ciencia, la inteligencia artificial, la deducción automática y el aprendizaje automático durante la segunda mitad del siglo XX. Esto permitiría que el progreso en estas diferentes áreas fuera aplicado para construir sistemas verdaderamente inteligentes, capaces de adquirir y derivar nuevo conocimiento.

2. Objetivos

El objetivo principal de este trabajo es el estudio formal de la utilidad y resultado de la síntesis de conceptos en términos de ganancia de información y refuerzo en sistemas de inferencia. Las medidas a desarrollar deben ser aplicables consistente y uniformemente tanto a la inferencia deductiva como a la inductiva.

El interés principal será la *evaluación* (y no la generación) de conceptos (hechos, reglas, teorías, teoremas, propiedades, etc.), aunque algún aspecto de generación pueda ser abordado puntualmente. Para la evaluación se deberán diseñar varias medidas coherentes y válidas tanto para la deducción como para la inducción. En concreto, los objetivos específicos de esta tesis vienen dados por la medición de las siguientes dimensiones:

- **Informatividad**: una nueva medida asignará la *ganancia de información* de una inferencia de un concepto x a un concepto y. Esto permitirá clarificar las nociones de explicitez e implicitez, y dar nociones generales y alternativas a la información superficial y profunda de Hintikka para sistemas deductivos y a la informatividad de Popper para la inducción.
- **Plausibilidad**: esta dimensión no es aplicable para la deducción clásica. Se medirá esta dimensión para procesos que no preservan la verdad (inducción, deducción aproximada o no monótona, abducción, ...) por medio de la teoría del *refuerzo*, dada por el uso necesario de cada parte de la teoría o sistema en el resto de la misma y en la evidencia.
- **Consiliencia**: esta dimensión, informalmente introducida por Whewell en 1847, se relaciona con el nivel de uniformidad con el que una teoría cubre sus consecuencias, y también se conoce (aunque con matices) como coherencia o unificación. Se ha vindicado en la inducción explicativa, donde la teoría debe ser comprehensiva con la evidencia, en el sentido de que todos los ejemplos deben ser cubiertos o unificados por la misma regla general (o *causa*).
- **Intensionalidad**: una cuestión primordial asociada con cualquier definición es si es extensional (por extensión) o intensional (por comprensión). Un primer análisis de la cuestión mostrará que no es apropiado asignar una respuesta booleana a la misma. Por consiguiente, se introducirá un grado de intensionalidad, estrechamente relacionado con la idea de excepción.
- **Comprensibilidad / Inteligibilidad**: Se introducirá una medida que escale la dificultad de comprender, concretamente el grado de comprensibilidad o inteligibilidad de un concepto cualquiera. También se particularizará para estimar la dificultad de diversos problemas de inferencia inductivos y deductivos. Esto permitirá la medida de capacidades intelectuales, sin contaminación antropomórfica.

- **Utilidad**: en sistemas deductivos, la utilidad de la introducción de nuevos conceptos para diferentes propósitos está clara de una manera informal: una mejor comprensión de la teoría en su conjunto, una expresión más concisa de la misma idea, una reducción del tiempo computacional y espacio de futuras deducciones, etc. Esta necesidad de información intermedia se mostrará formalmente. En el caso de la inducción, la noción de utilidad está estrechamente relacionada con la de plausibilidad, y vendrá representada por el uso del refuerzo para medir la utilidad en inferencia deductiva e inductiva.

Está claro que la mayoría de estas dimensiones son dependientes o contra-dependientes, y esto es admisible siempre y cuando éstas representen medidas útiles e intuitivamente diferentes. Este fenómeno está motivado también por el uso de diferentes mecanismos representacionales, y algunas de estas 'dependencias' no pueden elucidarse de una manera absoluta en cualquier mecanismo descripcional. En este sentido, la intención inicial es permitir que estas medidas puedan ser aplicadas a cualquier mecanismo de representación, aunque en algunos casos se asumirán algunas restricciones menores, con el fin de posibilitar medidas más finas y prácticas.

Se indagarán también medidas derivadas para la optimalidad de representación y el comportamiento global de un sistema axiomático. En el caso de la deducción, esto puede hacerse sin la contrastación con la experiencia exterior. Esta realimentación interna, similar a la noción de matemáticas experimentales, se muestra claramente en la teoría de juegos, un área que también fue abordada por Hintikka. Nociones fundamentales para entender la dinámica de estos sistemas, como la información intermedia y una noción genérica de simplificación, serán clarificadas.

Para el caso de la inducción, sin embargo, la evidencia exterior es el factor principal (pero no exclusivo) que determina la bondad de una teoría. No obstante, una teoría inductiva puede construirse por muy diferentes motivos: explicar la evidencia, predecir la evidencia futura, describir la evidencia, ser comprehensiva, etc.

Finalmente, la combinación de inducción, deducción, confirmación y ganancia de información se particularizará para teorías lógicas, pero también se ensayará su aplicación a diferentes aspectos de las bases de datos modernas y los sistemas software complejos, que comienzan a englobarse conjuntamente bajo el nombre de sistemas de conocimiento.

3. Discusión (Desarrollo)

Aunque este apartado se debe denominar "discusión"[4], es aquí donde se incluye el desarrollo de la tesis y la consecución de los objetivos, mientras que la sección siguiente ("conclusiones") es la que realmente discute los resultados de la tesis. La redacción sigue el orden del texto original de los capítulos 3 al 9, cada uno de ellos ocupando un subapartado de esta sección. No obstante, aparecen separados en dos partes: una primera parte donde se introducen las nuevas medidas y conceptos (así como algunas aplicaciones directas) desarrollados como consecución de la tesis (capítulos 3 al 6) y una segunda parte donde dichas medidas y nociones formales se aplican a muy diversos campos, mostrando su funcionamiento en la práctica y su verdadera utilidad (capítulos 7 al 9).

3.1 Nuevas Medidas y Conceptos Relacionados

Como hemos dicho, los siguientes cuatro subapartados corresponden a los capítulos 3 al 6. Las herramientas que se utilizan para su desarrollo son la complejidad de Kolmogorov, en especial, la variante de Levin espacio-temporal, y una sencilla teoría del refuerzo, aunque irá haciéndose más sofisticada a medida que avanza el desarrollo.

Aunque el texto original cuenta con el apéndice A que recoge los conceptos más importantes de la complejidad de Kolmogorov, incluimos aquí las definiciones estrictamente necesarias.

Denominamos una máquina universal ϕ a aquélla que puede emular una máquina universal de Turing. $\phi(p,y)$ denota el resultado de la ejecución de p en ϕ con entrada y. $Cost_\phi(p,y)$ denota el coste computacional (pasos de la máquina ϕ) de ejecutar el programa p con entrada y. A partir de aquí, se pueden dar las siguientes definiciones:

Definición 3.1 (*en la disertación 1.3*) **Complejidad de Kolmogorov**.

$$K(x|y) = \min \{ l(p) : \phi(p,y) = x \}$$

Además se exige que la máquina ϕ use una codificación para los programas que sea libre de prefijos.

El término x^* denota el primer programa mínimo para x según un orden de enumeración lexicográfico. Por tanto, $l(x^*) = K(x)$.

[4] Según la resolución de la Comisión de Doctorado de fecha 21 de junio de 1995.

Definición 3.2 (*en la disertación 1.4*) **Complejidad de Levin**.

$$Kt(x|y) = \min \{ LT_{\phi}(p) : \phi(p,y) = x \}$$

donde $LT_{\phi}(p) = l(p) + log_2 Cost_{\phi}(p,y)$.

Existen dos variantes absolutas (no relativas) de las complejidades anteriores, que se definen sencillamente como $K(x) = K(x \mid \varepsilon)$ y $Kt(x) = Kt(x \mid \varepsilon)$ donde ε es la cadena u objeto vacío.

Algunas indicaciones adicionales sobre terminología y notación usadas en la tesis se pueden encontrar al final del capítulo 1.

3.1.1 Nuevas Medidas de Ganancia de Información y Representación

El capítulo 3, *Information and Representation Gains*, introduce las primeras medidas (y las más teóricas) de la tesis. El propósito principal es evaluar la cantidad de información que se ha hecho explícita en un paso de razonamiento o de inferencia. Tras establecer unos criterios que una medida de este tipo debería tener y después de algunos intentos insatisfactorios (tanto de la literatura como del autor de esta tesis), y que se pueden encontrar al principio del capítulo, la primera medida útil se denomina *ganancia de información independiente del tiempo*. Esta nueva medida, denotada por $V(x|y)$, representa la porción de información de x que está implícitamente en y.

Definición 3.3 (*en la disertación 3.4*) La ganancia de información normalizada, relativa e independiente del tiempo de x con respecto a y, denotada por $V(x \mid y)$ se define como:

$$V(x \mid y) = K(x \mid y) / K(x)$$

La función $V(\cdot|\cdot)$ está indefinida sii $x = \varepsilon$ y está bien definida para cualquier $x \neq \varepsilon$ ya que $K(x) > 0$. Para todo x e y, es obvio que $1 \geq V(x \mid y) > 0$. El límite superior se da precisamente cuando y no contiene ninguna información sobre x, es decir, $I_c(x:y) = 0$. Esta situación significa que y es inútil para obtener x porque x e y son absolutamente independientes. El límite inferior se da cuando $x = y$ pero también cuando y es un programa para x.

Esta medida, aunque útil desde un punto de vista teórico para dar cuenta de la noción de implicitez, presenta algunos problemas para una medida de ganancia de información intuitiva. Aparentemente, si $V(x \mid y) \approx 0$, x se puede obtener de y, pero el tiempo computacional puede ser extremadamente alto, con lo que y en realidad no es útil para obtener x. En este sentido, V sólo recoge la idea de 'dificultad' o 'esfuerzo' como tamaño de los datos necesarios. Finalmente, el mayor problema es que $K(\cdot)$ no es computable, lo que lógicamente hace que V tampoco sea computable. Veamos por tanto una versión de la ganancia de información computable que, además, resuelve los problemas anteriores.

Como vimos en la introducción, en los sistemas no omniscientes, donde la noción de esfuerzo tiene sentido, la concepción intuitiva de información se reentiende en términos de consumo de recursos. La elección de la función LT, que pondera el espacio (datos extra) necesario y el tiempo, resuelve los problemas de considerar sólo el espacio (como hemos visto) o sólo el tiempo. A partir de la variante de Levin Kt definida a partir de LT, podemos definir una nueva función, denominada *ganancia de información computacional* $G(x|y)$, que depende del esfuerzo computacional (tiempo y espacio), y mide la proporción de x que puede ser obtenida fácilmente con la ayuda de y.

Definición 3.4 (*en la disertación 3.5*) La ganancia de información computacional (espacio-temporal) relativa y normalizada de x con respecto a y, denotada $G(x \mid y)$ se define como:

$$G(x \mid y) = Kt(x \mid y) / Kt(x)$$

De igual modo que antes, la función $G(x|y)$ no está definida sii $x = \varepsilon$ y está bien definida para cualquier $x \neq \varepsilon$ ya que $Kt(x) > 0$.

Algunas propiedades de G se estudian en el capítulo 3, por ejemplo sus límites:

Teorema 3.2 (*en la disertación 3.7*) Existe una constante c tal que para cada x e y, $\log l(x)/(l(x) + \log l(x) + c) < G(x \mid y) \leq 1$.

Por tanto, para cualquier x e y lo suficientemente grandes G puede estar entre casi 0 y 1. En el texto se introduce una variante, basada en máquinas de Turing transparentes que hace que los límites estén entre 0 y 1 para objetos x e y de cualquier tamaño.

Una propiedad más importante, que muestra la robustez de G, es la que establece el siguiente teorema:

Teorema 3.3 (*en la disertación 3.9*) Consideremos un algoritmo de aprendizaje A^* en $\mathcal{P}$ (i.e. de coste polinómico), en concreto $\exists p \in \mathbb{N}^+ : O(n^{p-1}) \leq O(A^*) \leq O(n^p)$, siendo A^* de tamaño constante, es decir $l(A^*) = c$. Este algoritmo transforma de manera determinista y en x, donde x es un programa para y, con $n = l(y)$. Existe un τ tal que para todo x e y, si $n > \tau$ y $Kt(x) > k \cdot p \cdot \log n$, entonces $G(x|y) \leq 2 / k$.

De una manera más sencilla, el teorema afirma que si x e y son suficientemente grandes y existe un algoritmo polinómico de y a x, entonces $G(x|y)$ debe ser baja. Esto significa que la medida de $Kt(x|y)$ es muy dependiente de la existencia de un algoritmo de y a x y de la complejidad que tenga dicho algoritmo. Además, muestra la diferencia entre deducción (se debe decir qué algoritmo o regla usar) y computación determinista. Esta diferencia viene representada por la longitud de A^*, que sí puede ser importante si el algoritmo no está predeterminado.

Finalmente, otra propiedad importante de G es su relación con el concepto de potencial algorítmico [Bennett 1988] (véase apéndice A), que es la formalización de la noción de concepto difícil de obtener. Formalmente, una cadena x es *k-potente* si k es el menor entero positivo tal que $Kt(x) \leq k \log l(x)$ [Bennett 1988].

Es de esperar que si x es potente no puede haber un y corto tal que $Kt(x \mid y)$ sea pequeño, porque esto implicaría que x podría obtenerse y describirse a partir de y y, por tanto, $Kt(x)$ sería también pequeño. El teorema siguiente formaliza y limita esta intuición:

Teorema 3.4 (*en la disertación 3.10*) Para todo x e y, si x es k-potente entonces $G(x \mid y) > 1 / k$.

Ya que el Teorema 3.4 marca un límite inferior, sirve para aquellos casos donde k es bajo, es decir cuando x es muy fácil de obtener, y entonces cualquier y es inútil para obtener x en menos esfuerzo (por ejemplo la cadena 1^n).

Finalmente, sería interesante distinguir cuándo la ganancia $G(x \mid y)$ se produce por la añadidura de información aleatoria y no relacionada o, por el contrario, se obtiene porque el esfuerzo computacional de y a x es alto, pero no se necesita información adicional. Como el resultado dado por $G(x \mid y)$ es relativo a la complejidad de x, este efecto se reduce de alguna manera en $G(x \mid y)$ pero no se elimina del todo. Es posible comparar $G(x \mid y)$ con $G(y \mid x)$ para excluir casos de información no relacionada. En general, sin embargo, es imposible saber *efectivamente* cuando x no contiene información aleatoria e independiente, porque esta información puede estar entrelazada con el resto de muchas maneras intrincadas (incluso criptográficamente). No obstante, $K(x \mid y)$ representa exactamente (que no computa) esta información común. Esto permite la siguiente definición:

Definición 3.5 (*en la disertación 3.7*) La ganancia real de información de x con respecto a y, denotada $TG(x \mid y)$ se define como:

$$TG(x \mid y) = (Kt(x \mid y) - K(x \mid y)) / Kt(x)$$

El nombre *ganancia real de información* se justifica por el hecho de que compensa lo que no está fácilmente obtenible en y o no está en y en absoluto ($Kt(x \mid y)$) y lo que no está en absoluto en y ($K(x \mid y)$). El resultado es exclusivamente una medida de lo que está en y pero no es fácil de obtener. En otras palabras, TG mide cuánta información de x está implícitamente en y.

En su artículo "*¿Cuándo se representa la información explícitamente?*" [Kirsh 1990], Kirsh afirma que la mayoría de las discusiones sobre conocimiento y representación caen en paradojas debido a nociones ambiguas de los términos 'explícito' e 'implícito'. Aunque movido inicialmente por la noción 'más profunda' de implicitez, pronto reconoce que la explicitez ha sido también muy problemática. Finalmente, con la premisa de que "*implícito es aquello que no es explícito pero podría serlo*" [Kirsh 1990], introduce una teoría informal de la explicitez, y pospone una teoría de la implicitez.

Pero como hemos visto, $V(x \mid y)$ y $G(x \mid y)$ generalizan y formalizan las nociones de implicitez y explicitez. La siguiente tabla resume esta relación:

$V(x \mid y)$	$G(x \mid y)$	$TG(x \mid y)$	$V(x \mid y)/G(x \mid y)$	*Significado*
1	1	0	1	x no está implícito ni explícito en y
0	1	1	0	x está profundamente implícito en y
1	$\cong 0$	-	-	*Imposible*
0	$\cong 0$	$\cong 0$	0	x está explícito en y

Tabla 3.1. *Diferentes casos y grados de implicitez y explicitez.*

Aunque $V(x \mid y)$ / $G(x \mid y)$ está bien definido y siempre entre 0 y 1, no es suficiente para separar los tres diferentes casos (filas) mostrados en la tabla 3.1, y TG tampoco es suficiente. Teóricamente, es necesario por tanto usar *dos* funciones, y en lo que sigue utilizaremos V y G. En definitiva, la tabla también muestra que sólo la explicitez es computable, por lo que los dos primeros casos de la tabla son indistinguibles efectivamente. Esto justifica el uso en la práctica de una única función $G(x \mid y)$ para ver cuánta información está explícita o implícita entre dos conceptos.

La noción de implícito y explícito establece la primera conexión con otros intentos de entender o esclarecer nociones relacionadas con la ganancia de información. Aunque en el capítulo 4 aparecen más conexiones con conceptos relacionados con procesos de inferencia, veamos ahora un concepto tan recurrente como ambiguo en la literatura: el de la información interesante o estética.

En este sentido, Nake [Nake 1974] sugirió que los datos que eran más interesantes y agradables estéticamente muestran un *ratio* ideal entre la información esperada y no esperada. Aunque la primera conexión de interés, estética y complejidad Kolmogorov (y en especial *Kt*) ha sido establecida siguiendo estas directrices, mediante la propuesta de un agente 'curioso' [Schmidhuber 1997b], esta idea se puede formalizar mediante la función G de la siguiente manera:

Definición 3.6 (*en la disertación 3.14*) **Interés**

Un concepto x es interesante para un agente con un conocimiento y iff

$$b - c < G(x \mid y) < b + c$$

donde $0 \leq b \leq 1$ es la audacia del agente y $0 \leq c \leq 1$ su umbral de curiosidad.

La función de ganancia de información permite comparar cualesquiera dos objetos x e y. En el caso en el que x es un programa o representación para y la ganancia puede todavía estar entre casi 0 y 1. No obstante, puede ser el caso que se quiera comparar tres objetos x, x' e y con las siguientes relaciones: x es una representación para y y x' es también una representación para y.

Esta cuestión es la que se estudia en la sección 3.8, que no discutiremos en detalle en este resumen, y que se dedica a comparar diferentes representaciones para un mismo objeto. La ganancia de representación se puede definir utilizando G

directamente entre x y x' (en los dos sentidos), cuando ambos son descripciones para y. Se plantean más casos que sería interesante formalizar, como qué representación es mejor que otra atendiendo a recursos computacionales, definiendo así un criterio de optimalidad de representación. Por este motivo también se estudian definiciones universales (y no semánticas) de simplificación, clarificando esta noción en sistemas deductivos. Finalmente se muestra que la ganancia de representación y la simplificación son nociones que se relacionan inversamente en general, como cabía esperar, ya que la ganancia incrementa la información y la simplificación generalmente la decrementa, mientras que la optimización busca un lugar intermedio.

En conclusión, algunos de los resultados de este capítulos son, *per se*, suficientemente importantes. Sin embargo, el potencial de estas definiciones, especialmente *G*, y su utilidad para aclarar cuestiones acerca de los procesos de inferencia, serán desvelados en el capítulo 4.

3.1.2 Ganancia de Información y Procesos de Inferencia

El capítulo 4, *Information Gain and Inference Processes*, saca partido de las definiciones y medidas introducidas en el capítulo anterior. Para ello, se debe basar en el hecho de que procesos aparentemente tan diferentes como son la inducción y la deducción pueden explicarse en un marco computacional como procesos de inferencia que generan una salida a partir de una entrada, y que deben cumplir ciertos criterios o restricciones, ampliamente estudiados en filosofía de la ciencia y en lógica matemática, respectivamente.

Por lo tanto es necesario clarificar y formalizar los diferentes tipos de sistemas deductivos con los que se trabajará y que son representativos de los sistemas de deducción tradicionales y automáticos:

> **Definición 3.7** (*en la disertación 4.15*) Un **Sistema Derivacional Computacional Determinista (DS)** se define como una máquina determinista ϕ que sólo acepta programas de la forma $\phi(<x, w>)$, donde x es un sistema axiomático y w es la información de selección que indica qué axiomas de x deben ser usados, qué ocurrencias deben seleccionarse y en qué orden. Finalmente, la siguiente condición se debe satisfacer:
>
> $\phi(<x, w>) = d \quad \rightarrow \quad x \vdash d$ y w es una demostración para d en x.

Por contra,

> **Definición 3.8** (*en la disertación 4.16*) Un **Demostrador de Teoremas Computacional (TP)** se define como una máquina determinista ϕ, que sólo acepta programas de la forma $\phi(<x, t>)$, donde x es un sistema axiomático y t es una fórmula bien formada de x tal que:
>
> $\phi(<x, t>) = 1w \quad \rightarrow \quad x \vdash t$, y w es una demostración para t en x y

$$\phi(<x, t>) = 0 \quad \rightarrow \quad x \not\vdash t$$

Se debe tener en cuenta que, para algunos sistemas axiomáticos altamente expresivos, es posible que ϕ no acabe.

Aunque la Definición 3.7 y la Definición 3.8 son 'equivalentes estructuralmente', la diferencia reside en la interpretación de la entrada (w y t respectivamente) y la salida (d y la demostración respectivamente).

Definición 3.9 (*en la disertación 4.17*) Un **Aceptador Computacional (AC)** se define como una máquina determinista ϕ, que sólo acepta programas de la forma $\phi(<x, t>)$, donde x es un sistema axiomático y t es una fórmula bien formada de x tal que:

$$\phi(<x, t>) = 1 \quad \rightarrow \quad x \vdash t, \text{ y}$$

$$\phi(<x, t>) = 0 \quad \rightarrow \quad x \not\vdash t$$

Del mismo modo se podrían definir los diferentes tipos de sistemas de aprendizaje o inductivos a partir de los sistemas deductivos anteriores, como los inductores clasificadores, inductores de modelo, etc.

En este marco computacional, la deducción y la inducción se pueden ver como mecanismos no omniscientes, es decir, procesos que consumen recursos.

Veamos en primer lugar el caso de la inducción. La ganancia de información computacional, es decir G, proporciona una medida uniforme del valor relativo o la informatividad de la hipótesis respecto a la evidencia (en el sentido de Popper), concretamente el esfuerzo computacional mínimo que ha de ser invertido desde los datos (evidencia) hasta la hipótesis. En particular, si x es la teoría e y es la evidencia, los dos casos extremos son claramente ilustrativos:

- Mínimo: $G(x \mid y) = \log l(x) \;/\; (l(x) + \log(l(x)) \approx 0$. La teoría es evidente a partir de los datos. En otras palabras, es muy fácil describir la teoría a partir de la evidencia. Algunos ejemplos de este mínimo son: una descripción llena de excepciones o con grandes extensionalidades, ya que pueden citarse fácilmente a partir de los datos, o el polinomio de orden n–1 obtenido a partir de n datos.
- Máximo: $G(x \mid y) = 1$. La teoría es sorprendente o creativa respecto a los datos. Los datos son inútiles (en términos espacio-temporales) para describir la teoría ($Kt(x \mid y) = Kt(x)$). Es necesario un trabajo computacional grande sobre los datos y para obtener la teoría y/o hay una necesidad de información externa. En otras palabras, el esfuerzo invertido justifica que x se conserve, porque es algo valioso.

Es importante resaltar que $G(x \mid y)$ mide la ganancia de y a x y no la plausibilidad de x. En otras palabras, la ganancia de información es un criterio puramente metodológico, porque no hay ninguna razón para pensar que las teorías 'complicadas' son más probables. Sin embargo, metodológicamente, G puede utilizarse para obtener un buen "criterio de olvido".

En particular, sería interesante mantener aquellas hipótesis que no han sido seleccionadas por el criterio de plausibilidad pero son explicaciones alternativas que podrían seleccionarse más tarde, si la evidencia futura descarta otras hipótesis momentáneamente mejores. El aprendiz ha invertido esfuerzo en obtener estas hipótesis y deben ser retenidas para recuperar una mayor parte de ese esfuerzo que sí sólo una fuera recordada. Formalicemos esta idea:

Definición 3.10 (*en la disertación 4.21*) **Criterio de Olvido.** Dado un criterio de plausibilidad $PC(h \mid d)$, y un aprendiz con hipótesis alternativas y recursos de memoria limitados, su política de memoria puede regirse por el siguiente criterio de olvido:

$$OC(h \mid d) = G(h \mid d) \cdot PC(h \mid d)$$

Las hipótesis con menor *OC* deberían olvidarse. Por ejemplo, si el criterio de plausibilidad fuera el principio MDL se tendría $OC(h \mid d) = G(h \mid d) \cdot 2^{-l(h)}$. Mediante el uso del factor *G*, por ejemplo, una hipótesis "print *d*" para unos datos aleatorios *d*, tendría $OC(h \mid d) = 0$, aunque el PC fuera el mayor. En definitiva, este criterio resulta ser un compromiso razonable entre informatividad y simplicidad. Para datos comprimibles, una hipótesis corta respecto a los datos es generalmente informativa (como veremos un poco más adelante) y el valor final de OC no se ve afectado demasiado por *G*. Por el contrario, para datos aleatorios, las hipótesis generadas por el principio MDL serían descartadas porque *G* es bajo[5].

Incluso este criterio podría utilizarse como criterio general, en relación con el dilema clásico entre hipótesis informativas y probables. Está claro que una explicación o teoría debe tener un grado de plausibilidad para evitar hipótesis fantásticas, pero en muchas aplicaciones, tales como el descubrimiento científico y la abducción, debemos considerar una explicación como una inversión, incluso una "apuesta arriesgada" que puede ser pronto falsificada. Éste no es más que el criterio de falsificabilidad de Popper [Popper 1962]: no siempre es deseable la explicación más probable, porque a veces es la menos falsificable / informativa también.

Este compromiso entre informatividad y plausibilidad también se encuentra en el Ratio de Ganancia de Quinlan, que es el centro del algoritmo ID3 de inducción de árboles de decisión [Quinlan 1986]. Su última implementación C4.5 [Quinlan 1993] es el programa más popular y más extensivamente usado de la comunidad de aprendizaje automático. Veamos su relación con la ganancia de información desde el punto de vista descripcional.

Siguiendo a [Quinlan 1993], si *C* es el conjunto de etiquetas de la clase, se puede obtener la entropía de *C* de la manera probabilística clásica (entropía de Shannon):

[5] Hay, por supuesto, otros aspectos que influyen sobre el criterio de olvido. En particular, la frecuencia de uso y el interés son, en muchos casos, más importantes que la plausibilidad o la ganancia. No obstante, como veremos, la frecuencia de uso, la plausibilidad y el interés (qué podría ser útil para el agente en el futuro) se incluyen en la medida de refuerzo que se presenta en el capítulo 5.

$$info(C) = H(C) = -\sum_{c \in C} P(c)\, log_2 P(c)$$

La entropía después de una partición del árbol resulta ser:

$$info_X(C) = \sum_{v \in X} P(v) \cdot H(C \mid v)$$

donde cada $H(C \mid v)$ es la entropía de cada subárbol que ha sido generado sabiendo v.

Es lógico pensar que lo que se ha ganado después de la partición es la diferencia en información entre la evidencia conjunta y la evidencia separada. Esto es lo que la ganancia de información clásica formaliza; la ganancia de considerar la característica X se mide mediante la diferencia en incertidumbre (i.e. entropía) entre las situaciones sin y con el conocimiento del valor de esa característica.

Definición 3.11 (*en la disertación 4.23*) **Ganancia de Información Clásica (o Probabilística) [Quinlan 1986]**

$$gain(X,C) = info(C) - info_X(C) = H(C) - \sum_{v \in X} P(v) \cdot H(C \mid v)$$

Se ha utilizado el término 'clásica' porque es simplemente una generalización de la ecuación tradicional de información dada por x sobre C cuando X tiene sólo un elemento x:

$$I(X : C) = H(C) - H(C \mid x)$$

donde $H(C \mid x)$ es la entropía condicional, definida como $H(C \mid x) = -\sum_{c \in C} P(c \mid x)\, log_2 P(c \mid x)$.

La ganancia de información, sin embargo, tiende a sobrestimar la relevancia de las características con un número grande de valores. En otras palabras, una característica biunívoca (que diera tantos subárboles como elementos) sería siempre exacta pero no generalizaría en absoluto, siendo completamente inútil. Para normalizar este valor para características con diferentes números de valores, [Quinlan 1993] introdujo una versión normalizada, llamada Ratio de Ganancia, como la Ganancia dividida por la Información de Partición *split info*(X), la entropía de los valores-característica.

Definición 3.12 (*en la disertación 4.24*) **Información de Partición [Quinlan 1993]**

$$split\ info(X) = -\sum_{v \in X} P(v) \cdot log_2 P(v)$$

Definición 3.13 (*en la disertación 4.25*) **Ratio de Ganancia [Quinlan 1993]**

$$gain\ ratio(X,C) = gain\ (X,C) \,/\, split\ info(X)$$

Veamos ahora la relación con los conceptos introducidos en esta tesis. Gracias a el teorema de igualdad (asintótica) entre la entropía estocástica de Shannon y la complejidad algorítmica o descripcional (demostrado por Kolmogorov, véase e.g. [Li and Vitányi 1997]), podemos realizar una traducción de las definiciones anteriores a complejidad descripcional:

Definición 3.14 (*en la disertación 4.26*) **Ganancia y Ratio de Ganancia Descripcionales**

$$\textit{desc gain}\ (X,C) = \textit{desc info}(C) - \textit{desc info}_X(C) = K(C) - K(C\,|X)$$

$$\textit{desc gain ratio}(X,C) = \textit{desc gain}\ (X,C) \,/\, \textit{desc info}(X) = \{\, K(C) - K(C\,|X)\,\} \,/\, K(X)$$

Curiosamente, estas definiciones se parecen a las de ganancia de información del capítulo 3. En concreto, se puede establecer la siguiente conexión:

Teorema 3.5 (*en la disertación 4.13*) Para cualquier X y C, *desc gain ratio*$(X,C) = 1 - V(X \mid C)$ hasta un valor aditivo independiente $c \leq c' \,/\, K(X)$, siendo c' otra constante independiente.

Ya que V está siempre entre 0 y 1, para valores grandes de X, *desc gain ratio* y V son simplemente *complementarios*. Si $V(X \mid C) \cong 1$, X es completamente nuevo o independiente para C, así que es inútil para hacer una partición en C, y *desc gain ratio*$(X,C) \cong 0$. Por el contrario si $V(X \mid C) \cong 0$, X está completamente incrustado en C, con lo que es muy útil para partir C, y, lógicamente, *desc gain ratio*$(X,C) \cong 1$.

La razón de dividir por $K(X)$ tanto en G y en *desc gain* es más bien la misma: no estamos interesados en un valor absoluto, porque si no, los X grandes darían siempre las mayores ganancias.

Existen muchos otros criterios para aprender árboles de decisión, por ejemplo el principio MDL, que se puede aplicar para construir el árbol que sea más corto de describir, y que da buenos resultados, al menos si los datos son comprmibles.

En este sentido, la visión de aprendizaje como compresión [Solomonoff 1964] también se sustenta por el hecho de que la compresión[6] y la informatividad están relacionadas positivamente en general. Veamos que es así, al menos para aprendices 'minuciosos':

Teorema 3.6 (*en la disertación 4.11*) Consideremos un aprendiz minucioso Λ, que examina todos los datos o nada de ellos. Si la hipótesis es muy corta respecto a la evidencia (el ratio de compresión $R(d{:}h) > l(d) \,/\, \log l(d)$) entonces $G(h \mid d) = 1$.

El objetivo del Teorema 3.6 juntamente con el Teorema 3.3 (en la disertación 3.9) es mostrar que los aprendices eficientes (computacionalmente hablando) no pueden obtener hipótesis informativas. En concreto, cuando se usa un algoritmo de aprendizaje de coste polinómico para aprender una evidencia de talla n, se necesita un ratio de compresión mucho mayor que $n \,/\, (p \log n)$, lo cual, para n grande y p pequeño (como suele ser el caso) es un requisito muy estricto y restrictivo. Esto apoya la tesis de que *los algoritmos eficientes que trabajan exclusivamente a partir de los datos no pueden aprender hipótesis valiosas*, o, visto del otro lado, los algoritmos eficientes siempre citan 'superficialmente' parte de los datos. Esto resalta la importancia del contexto, aquello que se viene dado, que sirve para generar las hipótesis mediante

[6] Hay que tener en cuenta que el principio MDL no asegura ninguna compresión en absoluto en general.

prueba y error (utilizando la realidad como oráculo), o por aproximaciones menos drásticas, tales como los algoritmos genéticos.

En este sentido, bajo la noción más intuitiva de aprendizaje, se considera que un concepto creativo no puede ser fácilmente obtenido a partir de algo que no se conocía anteriormente, porque en este caso no sería novedoso. Por tanto, la clave de la creatividad, si la hay, se puede encontrar en la repetición de estructuras y patrones previamente existentes, con el fin de hacer $Kt(x \mid b) = Kt(x)$, donde x representa el concepto 'novedoso' y b es el conocimiento previo.

Una reflexión de los resultados anteriores para el campo del aprendizaje automático sugiere que si la hipótesis es evidente a partir de los datos, no se debe haber producido demasiado aprendizaje. Esto se muestra todavía con mayor rotundidad cuando los datos son aleatorios (y esto sucede habitualmente cuando los datos son cortos porque no merece la pena comprimirlos). Por ejemplo, el principio MDL daría los datos mismos, lo cual no corresponde con la idea de 'modelo' o explicación. No obstante, los paradigmas de aprendizaje más importantes se basan en la idea de la identificación: identificación en el límite [Gold 1967], el modelo PAC [Valiant 1984], Aprendizaje por Consultas [Angluin 1988]. Estos paradigmas están diseñados para datos infinitos, pero un algoritmo de aprendizaje que siempre dé una descripción completamente extensional (y no valiosa) "print x" para cualquier evidencia finita x ha identificado lo datos, y habría, en términos formales, aprendido, lo cual es bastante contra-intuitiva.

Por esta razón, y muy lejos de la noción clásica de 'identificación' [Gold 1967], propongo una noción diferente de aprendizaje (o descubrimiento): cuanto más valiosa es la descripción en relación a los datos, más aprende el sistema[7].

Definición 3.15 (*en la disertación 4.22*) **Aprendizaje Auténtico o Descubrimiento**. Decimos que un concepto o teoría x es un *aprendizaje auténtico* o *descubrimiento* con respecto a y en un contexto β si x es una teoría o descripción para y y $G(x \mid \langle y, \beta \rangle)$ es cercano a 1.

Con esta definición dejamos de momento la aplicación de G a la inducción y pasamos a ver cómo se comporta con la deducción.

En el caso de la deducción, $G(x \mid y)$ también proporciona una medida uniforme del valor relativo de las conclusiones con respecto a las premisas, la ganancia del esfuerzo computacional que se ha invertido en el proceso de las premisas a las conclusiones. Más concretamente, si x es la conclusión e y representa las premisas, los dos casos extremos siguientes son ilustrativos:

- Mínimo: $G(x \mid y) = \log l(x) \,/\, (l(x) + \log(l(x))) \approx 0$. La conclusión es evidente a partir de las premisas. En otras palabras, es muy fácil describir la conclusión a partir de las premisas. Algunos ejemplos que pueden producir este mínimo

[7] La noción de aprendizaje auténtico (vista como saber algo que no se sabía implícitamente) se puede aplicar a la deducción, como ahora veremos, y daría lugar a que no sólo se puede aprender por medios inductivos, sino que el aprendizaje auténtico también se puede producir por medio de la inferencia deductiva.

son: una conclusión que añade alguna tautología sencilla a las premisas, una conclusión que cambia el orden de algunos componentes lógicos, una conclusión que es una instancia directa de una premisa o una conclusión compuesta mayoritariamente de las premisas y algunas pocas cosas derivadas.

- Máximo: $G(x \mid y) = 1$. La conclusión es sorprendentemente o incluso creativa respecto a las premisas. Las premisas son finalmente inútiles (en términos espacio-temporales) para describir la conclusión ($Kt(x \mid y) = Kt(x)$). Se necesita un trabajo computacional grande sobre las premisas y para obtener la conclusión x o se necesita información externa. En otras palabras, el esfuerzo computacional invertido justifica que x se mantenga. Cualquier teorema difícil es una muestra de esto.

La mayoría de las deducciones caen entre estos dos casos extremos. Es interesante comparar este análisis con el mismo que se hizo sobre el uso de *G* para la inducción.

Para obtener una interpretación más detallada, se aplican las funciones *V* y *G* para los diferentes paradigmas deductivos que se han visto anteriormente, obteniendo como resultado las siguientes aproximaciones que se muestran en la Figura 3.1 (*en la disertación 4.2*)

Tipo de Sistema Deductivo	*V*	*G*
Sistema Derivacional		
• Sin demostración: $F_\phi(d \mid x)$	$\leq \min_{w\,:\,\phi(<x,w>)=d} K(w) \,/\, K(d)$	$\leq \min_{w\,:\,\phi(<x,w>)=\,d} \{\, Kt(w) + \log \text{Cost}(\phi(<x,w>)) \,\} \,/\, Kt(d)$
• Con demostración: $F_\phi(d \mid <x,w>)$	$= 0$	$\leq \min\,(\log \text{Cost}(\phi(<x,w>)) \,/\, Kt(d), G_\phi(d, x))$
Demostrador: $F_\phi(w \mid <x,t>)$	$= 0$	*bastante variable*
Aceptador: $F_\phi(a \mid <x, t>)$	$= 0$	1
Clasificador: $F_\phi(a \mid <x, t>)$	$= 0$ si n grande	*bastante variable.*

Figura 3.1. *Diferentes aproximaciones a V y G para diferentes sistemas deductivos.*

Del mismo modo, y para poder aplicarlas en la práctica, se derivan aproximaciones para teorías lógicas, utilizando estimaciones del coste descripcional para teorías lógicas, como *L* y L_{PC}, tal como se muestra en la Figura 3.2 (*en la disertación 4.3*).

Tipo de Sistema Deductivo de 1er Orden	V	G
Sistema Derivacional		
• Sin demostración: $F_\phi(E \mid T)$	$\leq L(E\|T)/L(E)$	$\leq \{L(E\|T)+\log Cost(E\|T)\} / (L(E)+\log CostPrint(E))$
• Con demostración: $F_\phi(E\|<T, W>)$	$= 0$	$\leq \log Cost(E \mid <T, W>) / (L(E) + \log CostPrint(E))$
Demostrador		
• Dem. Canónica: $F_\phi(W_c\| <T, E>)$	$= 0$	$\leq \log Cost(E\|T) / (L(W_c) + \log CostPrint(W_c))$
• Otra demostración: $F_\phi(W\|<T,E>)$	$\leq L_{PC}(W\|<T,E>) / L(W)$	$\leq \{ L_{PC}(W\|<T,E>) + \log Cost(E \mid <T, W>) \} / (L(W) + \log CostPrint(W))$
Aceptador: $F_\phi(a \mid <T, E>)$	$= 0$	$\leq \log Cost(E\|T) /(n+\log n)$

Figura 3.2. *Diferentes aproximaciones para V y G para diferentes sistemas de 1er orden.*

Desde el punto de vista teórico, y volviendo a representaciones universales, G representa una formalización de las ideas de Hintikka independiente del número de individuos y del lenguaje descripcional usado. Además, $G(\cdot|\cdot)$ es computable y tiene en cuenta el tiempo, diferenciando claramente lo que está disponible fácilmente en un momento dado de aquella información intrincada que requiere esfuerzo de extraer. De todos modos, sería interesante establecer la conexión de los conceptos generales de G y V con los conceptos de Hintikka de información superficial e información profunda, respectivamente.

Hintikka mostró que la información profunda es el límite de la información superficial. Podemos establecer un paralelismo similar con G y V. Consideremos la notación $speedup(\phi_a, \phi_b) = n$ para denotar que ϕ_a es n veces más rápido que ϕ_b, i.e., para cada paso que ϕ_b realiza, ϕ_a realiza al menos n en el mismo tiempo, o, alternativamente, que ϕ_a hace en 1 paso al menos n operaciones de ϕ_b. A partir de aquí:

Teorema 3.7 (*en la disertación 4.14*) Seleccionemos cualquier máquina universal ϕ, y definamos G_n en una máquina ϕ_n tal que $speedup(\phi_n, \phi) = n$ y que para cualquier programa efectivo p se tiene que $\phi_n(p)=\phi(p)$. Entonces para cualquier par de conceptos finitos x,y se tiene que $\lim_{n\to\infty} G_n(x \mid y) = V_\phi(x \mid y)$.

Finalmente, en el apartado 4.7 del capítulo se retoman las nociones de optimización representacional del capítulo anterior, y se encuentra un compromiso entre el tamaño y el tiempo requerido para obtener la evidencia a partir de la teoría. Se particulariza a

diferentes paradigmas deductivos, mostrando el papel de la información intermedia en el campo de la demostración automática de teoremas y en la práctica matemática. También se introduce una extensión descripcional de la noción de Poder Sistemático de Pietarinen, dando de nuevo las expresiones $1 - V(d \mid h)$ y $1 - G(d \mid h)$, que corresponde con la 'satisfactoriedad' potencial relativa de Popper.

En el apartado 4.8 se formaliza la diferencia entre razonamiento anticipado (eager) y perezoso (lazy), considerando el tiempo en el que se realiza el esfuerzo computacional y/o cómo se usan los recursos espaciales.

Por último, el apartado 4.9 aborda la relación entre inducción, deducción e información, sin olvidar el aspecto de la plausibilidad. Se ve cómo los criterios inductivos existentes son incompatibles con la información intermedia de los sistemas deductivos. La discusión concluye con la convicción de que sólo es posible conciliar inducción, deducción e información si se tiene en cuenta los recursos y se evita la omnisciencia, e incluso es conveniente contemplar la posibilidad de inconsistencias en el conocimiento. La figura 4.4 de la disertación muestra también las combinaciones de los axiomas modales T, K y D y su influencia en los procesos de inferencia, cuándo son necesarios y, por tanto, cuándo hace falta utilizar V y G. Finalmente también se reconoce que una medida de utilidad o plausibilidad es necesaria para dar cuenta del valor total de un proceso de inferencia.

3.1.3 Evaluación mediante Refuerzo Constructivo

El capítulo 5, *Reinforcement,* presenta una medida operativa de confirmación para teorías constructivas generales, estudiando la construcción de conocimiento, la revisión, la abducción y la deducción en este mismo marco.

Sea cual sea la aproximación a la construcción de conocimiento, la revisión del conocimiento se puede deber ya sea a una inconsistencia o a una falta de soporte. En este último caso, una debilidad parcial o total de la teoría se puede detectar por una falta de *refuerzo* (o repartición del crédito [Holland et al. 1986]). Han habido varias justificaciones empíricas y teóricas al refuerzo en diferentes campos, desde las observaciones empíricas de los procesos de aprendizaje en los animales y humanos hasta las verificaciones teóricas y prácticas mediante validación cruzada.

El estudio del aprendizaje por refuerzo como técnica ha sido especialmente fructífero en esta última década (véase [Kaelbling et al. 1996]). Uno de los mayores problemas del aprendizaje por refuerzo es que es tanto más difícil asignar y 'propagar' el refuerzo dependiendo de dos factores (que también están relacionados): (1) cuán anticipativa es la estrategia inductiva (frente a métodos perezosos) y (2) cuán expresivo es el lenguaje donde la inducción se da lugar. Por tanto, el razonamiento basado en explicaciones (EBL) y la programación lógica inductiva (ILP) son dos áreas donde sería muy útil poder aplicar la teoría del refuerzo.

Es precisamente en lenguajes expresivos (universales) y anticipativos (que crean un modelo o teoría) donde se va a desarrollar la teoría. Además se abordará y se

resolverá el problema del refuerzo para lenguajes constructivos. Un lenguaje constructivo es aquél que permite crear nuevos términos (predicados o funciones) para expresar la evidencia de una manera más compacta. Precisamente, esto permite la introducción de conceptos fantásticos, problema que se resolverá por primera vez en la literatura del refuerzo.

Vamos a estudiar la propagación del refuerzo para lenguajes que estén formados de reglas o componentes *r* que se compongan de una cabeza (o consecuencia) y un cuerpo (o conjunto de condiciones). Cada regla se denota de la siguiente manera $r \equiv \{ h :- t_1, t_2, ... t_s \}$. Una teoría es simplemente un conjunto de reglas: $T = \{r_1, r_2, ..., r_m\}$. Esto permite particularizar todo lo que se desarrolle a continuación para lenguajes proposicionales, teorías Horn, lenguajes funcionales, gramáticas y muchos lenguajes de orden superior. Más aún, en lo que sigue, la semántica se dejará sin especificar y diremos simplemente que *e* es una consecuencia de *P*, denotado $P \models e$ (en otras palabras, hay una demostración para *e* en *P*, o, sencillamente, *P* cubre *e*).

A partir de aquí, comenzamos a definir algunas construcciones básicas:

Definición 3.16 (*en la disertación 5.35*) Diremos que una regla r_i es necesaria con respecto a *T* para un ejemplo *e* sii

$$T \models e \quad \text{y} \quad T - \{r_i\} \not\models e$$

A partir de aquí,

Definición 3.17 (*en la disertación 5.36*) Una teoría *T* está reducida para un ejemplo *e* sii

$$T \models e \quad \text{y} \quad \neg\exists\, r_i \in T \text{ tal que } r_i \text{ no es necesaria para } e$$

Consideraremos una demostración como un conjunto de reglas, independientemente de su orden de combinación, las sustituciones aplicadas o el número de veces que cada regla se use. Esta concepción inusual (e incompleta) de demostración nos permitirá trabajar sin considerar una semántica concreta y mantener un grado apropiado de detalle. Gracias a esto la siguiente definición es posible:

Definición 3.18 (*en la disertación 5.37*) Diremos que S_1 y S_2 son demostraciones alternativas para un ejemplo *e* en una teoría *T* sii

$$S_1 \subset T,\ S_2 \subset T,\ S_1 \neq S_2 \text{ y } S_1 \text{ y } S_2 \text{ están reducidas para } e$$

Denotaremos por $Proof(e,T)$ el conjunto de demostraciones alternativas para un ejemplo *e* con respecto a una teoría *T*. Finalmente, definimos $Proof_r(e,T)$ como el conjunto de demostraciones alternativas que contienen *r* . Más formalmente,

Definición 3.19 (*en la disertación 5.38*)

$$Proof_r(e,T) = \{ S : S \subset Proof(e,T) \text{ y } r \in S \}$$

Con estas sencillas construcciones, podemos introducir ya la primera medida de refuerzo:

Definición 3.20 (*en la disertación 5.39*) El refuerzo puro $\rho\rho(r)$ de una regla r de una teoría T con respecto a una evidencia dada $E = \{e_1, e_2, ..., e_n\}$ se define como:

$$\rho\rho(r) = \Sigma_{i=1..n}\, card(Proof_r(e_i,T))$$

En otras palabras, $\rho\rho(r)$ se computa como el número de demostraciones de e_i donde se usa r. Si hay más de una demostración para un mismo e_i, *todas* ellas se tienen en cuenta, pero en la misma demostración, una regla sólo se computa una vez.

Definición 3.21 (*en la disertación 5.40*) El refuerzo (normalizado) se define como:

$$\rho(r) = 1 - 2^{-\rho\rho(r)}.$$

Esta definición hace que el refuerzo esté entre 0 y 1.

La media de las reglas de una teoría se define simplemente como:

Definición 3.22 (*en la disertación 5.41*) El refuerzo medio $m\rho(T)$ se define así:

$$m\rho(T) = \Sigma_{r\in T}\, \rho(r)/m,$$

siendo m el número de reglas.

A partir de estas definiciones se puede comprobar que, en general, la teoría más reforzada (en media) no es la más corta, como muestra el siguiente ejemplo:

Ejemplo 3.1 (*en la disertación 5.1*)

Dada una evidencia e_1, e_2, e_3, consideremos una teoría $T_a = \{r_1, r_2, r_3\}$ donde $\{r_1\}$ cubre $\{e_1\}$, $\{r_2\}$ cubre $\{e_2\}$ y $\{r_3\}$ cubre $\{e_3\}$ y una teoría $T_b = \{r_1, r_2, r_3, r_4\}$ donde $\{r_1, r_4\}$ cubre $\{e_1\}$, $\{ r_2, r_4 \}$ cubre $\{e_2\}$ y $\{r_3, r_4 \}$ cubre $\{e_3\}$.

A partir de las definiciones anteriores, T_a está menos reforzada que T_b.

En el primer caso tenemos $\rho\rho_{a,1} = \rho\rho_{a,2} = \rho\rho_{a,3} = 1$ y $m\rho(T_a) = 0.5$. Para T_b tenemos $\rho\rho_{b,1} = \rho\rho_{b,2} = \rho\rho_{b,3} = 1$, $\rho\rho_{b,4} = 3$ y $m\rho(T_b) = 0.5938$.

Además, la redundancia tampoco implica una pérdida de refuerzo medio (e.g. simplemente añade dos veces la misma regla).

No obstante, medir el refuerzo de la *teoría* presenta problemas de conceptos *fantásticos* (irreales), como muestra el siguiente teorema:

Teorema 3.8 (*en la disertación 5.18*) Consideremos un programa P compuesto de reglas r_i de la forma $\{ h :- t_1, t_2, .. t_s \}$, el cual cubre n ejemplos $E = \{ e_1, e_2, ... e_n \}$. Si el refuerzo medio $m\rho < 1 - 2^{-n}$ entonces siempre se puede incrementar mediante el uso de un concepto fantástico.

A primera vista, se podría deducir de este problema que el refuerzo debe combinarse con otro criterio (como el de simplicidad) para hacerlo funcionar (de hecho nunca se ha aplicado a lenguajes constructivos debido a este problema). Sin embargo, existe una solución sin acudir a otro criterio. La idea es medir la validación con respecto a la *evidencia*.

Definición 3.23 (*en la disertación 5.42*) El curso $\chi_T(f)$ de un hecho f con respecto a una teoría T se define como:

$$\chi_T(f) = max_{S \subset Proof(f, T)} \{ \Pi_{r \in S} \rho(r) \}$$

Más constructivamente, $\chi_T(f)$ se obtiene como el producto de todos los refuerzos $\rho(r)$ de todas las reglas r de T que se usan en la demostración de f. Si una regla se usa más de una vez, se computa más de una vez. Si f tiene más de una demostración, se selecciona el curso mayor.

Ahora sí el siguiente teorema muestra que no es posible incrementar el refuerzo mediante conceptos fantásticos:

Teorema 3.9 (*en la disertación 5.19*) El curso de cualquier ejemplo no puede incrementarse por medio de conceptos *fantásticos*.

Una vez hemos introducido una medida de refuerzo robusta, vayamos a aplicarla a evaluar teorías. La primera idea es utilizar la mayor media de todos los cursos de todos los ejemplos de la evidencia presentados hasta un momento dado:

Definición 3.24 (*en la disertación 5.43*) El curso medio $m\chi(T, E)$ de una teoría T con respecto a una evidencia E se define como:

$$m\chi(T, E) = \Sigma_{e \in E} \chi_T(e)/n$$

siendo $n = card(E)$.

Para obtener una teoría más compensada, se puede utilizar una media geométrica en vez de la aritmética, que denotaremos $\mu\chi$. Para cada teoría T, diremos que es *valiosa* para E sii $m\chi(T, E) \geq 0.5$, que es el caso límite cuando cada regla cubre sólo un ejemplo. Si el lenguaje de representación es lo suficientemente expresivo, es fácil mostrar que para cualquier evidencia E existe al menos una teoría valiosa para ella (simplemente elige la teoría con una regla extensional que cubra cada ejemplo). Lo mismo se cumple para $\mu\chi$.

El uso de esta medida puede verse en el ejemplo 5.2 de la disertación como un nuevo criterio válido para la construcción y revisión de conocimiento, funcionando tanto para la inducción como para la abducción.

Además, algunas nociones informales se pueden definir con la ayuda de la teoría del refuerzo. Por ejemplo, se puede formalizar la idea de 'consiliencia', introducida por Whewell en el siglo XIX [Whewell 1847], y otros conceptos relacionados, como el principio de Reichenbach de la causa común y el de coherencia de Thagard [Thagard 1978]. Todos ellos comparten la idea común de dar un teoría conciliadora para todos los ejemplos, es decir, toda la evidencia debe darse cuenta por la misma explicación o por explicaciones muy relacionadas.

Veamos que, bajo el contexto de refuerzo, se puede definir *consiliencia*:

Definición 3.25 (*en la disertación 5.44*) Una teoría T es particionable con respecto a una evidencia E sii $\exists T_1, T_2 : T_1 \subset T, T_2 \subset T$ y $T_1 \neq T_2$ tal que $\forall e \in E : T_1 \models e \vee T_2 \models e$. Definimos $E_1 = \{ e \in E : T_1 \models e \}$, $E_2 = \{ e \in E : T_2 \models e \}$ y $E_{12} = E_1 \cap E_2$. Finalmente, utilizaremos el término $S\chi(T_1 \oplus T_2, E)$ para denotar la expresión $m\chi(T_1, E_1) \cdot [\, card(E_1) - card(E_{12})/2 \,] + m\chi(T_2, E_2) \cdot [\, card(E_2) - card(E_{12})/2 \,]$.

Definición 3.26 (*en la disertación 5.45*) Una teoría T es consiliente con respecto a una evidencia E sii no existe una partición T_1, T_2 tal que $S\chi(T_1 \oplus T_2, E) \geq m\chi(T, E) \cdot card(E)$.

Aparte de ser directamente aplicable a cualquier teoría basada en reglas, veamos como se relaciona con la noción de excepciones o parches extensionales a una teoría. Antes de ello veamos que mediante el uso de refuerzo es fácil definir una excepción intrínseca como una regla r con $\rho = 0.5$, es decir, una regla que sólo cubre un ejemplo e. Pero debemos distinguir entre una excepción extensional completa, que se da cuando r no usa ninguna regla de la teoría para cubrir e, y excepciones extensionales parciales cuando r utiliza otras reglas para describir e.

La relación entre extensionalidad y refuerzo se muestra en el siguiente teorema:

Teorema 3.10 (*en la disertación 5.20*) Si una teoría T para una evidencia E tiene una regla r con $\rho = 0.5$, y completamente extensional, entonces T no es consiliente.

También vamos a establecer la conexión entre intensionalidad (vista como la prohibición o inexistencia de partes extensionales o excepciones, tal y como se acaban de definir) y el método de validación cruzada.

El tipo de validación cruzada que utilizaremos es la validación cruzada de particiones múltiples, que tiene en cuenta todas las posibles particiones en todas las posibles ordenaciones. Denotemos con n_e el número de reglas r que cubren un solo ejemplo e. En otras palabras, las reglas en las que si su ejemplo no apareciera, serían inútiles. Haremos además la siguiente suposición razonable: un algoritmo natural es un algoritmo que no añade reglas innecesarias a la teoría.

Definamos $P(A,T,E,k)$ como la probabilidad de que el algoritmo A dé la teoría T con los primeros k ejemplos de la evidencia E, considerando todas las posibles ordenaciones de E.

Teorema 3.11 (*en la disertación 5.21*) Para cualquier algoritmo de aprendizaje natural A,

$$P(A,T,E,k) \leq 1 - [(n-n_e)^{n-k} \,/\, n^{n-k}]$$

siendo $n = \text{card}(E)$.

El resultado puede entenderse como que uno debería evitar excepciones, con el fin de que $P(A,T,E,k)$ se acercara a 1. Por ejemplo, dada un teoría con 3 reglas de excepción (parches) para una evidencia de 100 ejemplos, tenemos que la probabilidad de que la teoría se encuentre con ochenta ejemplos es $P(A,T,E,k) \leq 1 - 97^{20}/100^{20} = 0.46$.

También se establece la relación positiva entre procesos de inferencia analógicos y la noción de consiliencia:

Teorema 3.12 (*en la disertación 5.22*) Si b implica E_1, c implica E_2, b no implica E_1 y c no implica E_2, la nueva teoría $T = \{ b', c', a \}$ tal que $T_1 = \{ b', a \} \models E_1$ y $T_2 = \{c', a\} \models E_2$, y ningún otro conjunto propio de T cubre ningún ejemplo, es consiliente.

Todavía existe un detalle que resolver con la medida de refuerzo que estamos utilizando. Existe una argucia para incrementar el refuerzo: unir reglas. Si un lenguaje de representación permite reglas muy expresivas, se podría hacer una teoría de tal manera que todo estuviera contenido en una única regla, que sería consiliente y reforzada por toda la evidencia.

Con el objetivo de mantener la granularidad de la teoría se opta por introducir un factor a cada regla inversamente proporcional a su longitud. Mediante $\text{length}(r)$ denotaremos la longitud de una regla r para un lenguaje en concreto. La única restricción es que, para todo r, $\text{length}(r) \geq 1$. Podemos ya extender las definiciones:

Definición 3.27 (*en la disertación 5.48*) El refuerzo puro extendido se define como:

$$\rho\rho^*(r) = \rho\rho(r) / \text{length}(r).$$

El refuerzo normalizado extendido $\rho^*(r)$ y el curso extendido $\chi^*(r)$ se definen de manera obvia. Esta modificación, no obstante, todavía hace muy diferente esta medida del principio MDL. Además, todos los resultados anteriores se siguen cumpliendo bajo esta extensión. Es fácil demostrar—en el límite— que la compresión es un excelente principio para obtener altos refuerzos:

Teorema 3.13 (*en la disertación 5.23*) Si los datos E son infinitos y la teoría T es finita, el curso medio $m\chi^*(T, E) = 1$.

La teoría del refuerzo también se puede extender para contemplar refuerzo negativo (se estudian varias opciones $\chi^1(f)$, $\chi^2(f)$ y $\chi^0(f)$), como se puede ver en las secciones 5.8 y 5.9, reconectándose con nociones más clásicas del refuerzo, basadas en recompensas y castigos.

Finalmente, el estudio del refuerzo para la deducción, muestra que también sirve como un criterio de utilidad, para saber si una propiedad, un lema o un teorema es útil para el resto de la teoría. Además, en el caso de la deducción, la confirmación se propaga completamente y para ello se establecen dos medidas alternativas de transmisión de plausibilidad por medio de la deducción de la siguiente manera:

Definición 3.28 (*en la disertación 5.59*) Para cada regla s tal que $T \models s$ se define su plausibilidad como:

$$P_1(s) = max_{S \subset Proof(s,T)} \{ \Pi_{r \in S} \rho(r) \}$$

O alternativamente,

Definición 3.29 (*en la disertación 5.60*) Para cada regla s tal que $T \models s$ se define su plausibilidad como:

$$P_2(s) = max_{S \subset Proof(s,T)} \{ min_{r \in S} \rho(r) \}$$

La Definición 3.29 concuerda con muchos trabajos sobre lógicas con valores de incertidumbre. Por ejemplo, si C representa el grado de certeza de un hecho cualquiera, $C(p \wedge q) = \min (C(p), C(q))$ y $C(p \vee q) = \max (C(p), C(q))$. Por el contrario, la Definición 3.28 se parece a otras teoría populares de la incertidumbre en las que C se mide entre 0 y 1, $C(p \wedge q) = C(p) \cdot C(q)$ y $C(p \vee q) = 1 - (1 - C(p)) \cdot (1 - C(q))$.

Estas dos funciones P_1 y P_2, junto con ρ y χ, pueden usarse como una medida cuantitativa de confirmación, de alguna manera entre Carnap y Hempel. En particular, son compatibles con muchas de las condiciones de idoneidad de Hempel, pero de una manera cuantitativa, como se discute en la sección 5.11. Ésta es la única manera, en mi opinión, de incluir tanto H2 (hacia abajo) y H5 (hacia arriba). Como resultado, la teoría del refuerzo es también una teoría entre el principio MDL y la informatividad de Popper, que resulta ser válida para la deducción, la inducción, la analogía y la abducción.

Por último, se compara la teoría del refuerzo con la ganancia de información. El criterio de olvido presentado anteriormente se particulariza con el uso del refuerzo:

$$OC(h \mid d) = G(h \mid d) \cdot m\chi(T, E)$$

Este criterio se utiliza para determinan si una regla se debe mantener o no en una teoría, ya se haya obtenido por inducción o por deducción.

El último problema a abordar, y aparcado hasta el final, es el cómputo de los valores de refuerzo. Se presenta un algoritmo incremental para ello y su complejidad se examina en la sección 5.13, con la convicción de que, aunque implica un coste adicional para cualquier algoritmo inductivo, se contrarresta por las ventajas de su uso como guía para la revisión de teorías, prorrateo de hipótesis, gestión de la evidencia y la memoria, etc.

En definitiva, uno de los resultados más importantes de este capítulo 5 es que la manera en la que el refuerzo se distribuye a través de la teoría resulta en una ontología *escalonada* que permite evaluar la teoría en su conjunto o parcialmente. De esta manera, uno de los dilemas más difíciles del aprendizaje inductivo, la elección de una distribución a priori, desaparece. En otras palabras, no es necesario trabajar con probabilidades para saber la plausibilidad total y detallada de cada regla de la teoría y de cada hecho o propiedad que se deriva de ella.

3.1.4 Intensionalidad y Explicación

El capítulo 6, *Intensionality and Explanation,* aborda el problema de distinguir formalmente entre una definición o descripción extensional y una intensional (o por

comprensión). En concreto, sería interesante distinguir aquellas descripciones que cumplen el siguiente requerimiento de comprensión: "*lo definido no puede aparecer en la definición*" [Bochenski 1965]. Este eslogan se sigue firmemente en los diccionarios y se utiliza por los maestros cuando preguntan a sus alumnos, con el fin de saber si han comprendido un concepto.

Adicionalmente, el uso tradicional en matemáticas distingue una definición extensional de un definición intensional (o por comprensión). Sin embargo, esta distinción es completamente intuitiva y ha existido un interés escaso por formalizarla, porque, para conjuntos infinitos, frecuentes en matemáticas, todas las definiciones deben ser intensionales (o por comprensión). No obstante, para conjuntos *finitos*, no se ha presentado hasta la fecha una diferencia formal entre una descripción intensional y una extensional. Esta noción es bastante difícil de atrapar formalmente para conceptos finitos porque hay muchas maneras diferentes de *disfrazar* una descripción extensional para parecer una descripción intensional.

Una primera formalización para el caso de teorías lógicas se puede realizar por medio de la noción de partición y evitando las excepciones, éstas vistas como partes extensionales o no validadas de una teoría, de modo similar al capítulo 5. Aunque algunas de las variantes presentadas se pueden utilizar en la práctica, el resto del capítulo se dedica a extender la idea para cualquier tipo de lenguaje.

La primera aproximación en este sentido se basa en la idea de que "*una excepción es algo que se puede quitar de una descripción, dejándola mucho más simple con respecto a la magnitud de la evidencia eliminada o no cubierta*". Más concretamente, una descripción está libre de excepciones si no existe una subdescripción que produzca casi todos los datos, es decir, no hay una reducción en la descripción que pueda ser mayor que la reducción correspondiente en los datos descritos. Formalicemos esto:

Definición 3.30 (*en la disertación 6.69*) Una descripción p_x para los datos x está c-libre de excepciones (denotado $\Delta_c(p_x) = 0$) sii no existe un subprograma p_y de p_x, siendo p_y un programa para y con $y \subset x$, tal que $K(p_x) - K(p_y) \geq [K(x) - K(y)] \,/\, c$. Nótese que en el caso de que exista, $p_x - p_y$ es la excepción (y p_y la regla general).

El parámetro c se puede estimar dependiendo del marco deductivo y la aproximación usada para calcular K, que podría ser Kt o, en su defecto, simplemente la función de longitud.

La definición anterior de descripción libre de excepciones es lo suficientemente general para adaptarse a cualquier lenguaje descriptivo. Sin embargo, esta generalidad hace difícil la comparación con otras nociones relacionadas y no se puede hacer operativa a no ser que el lenguaje descriptivo permita una noción de subprograma (como las diferentes versiones de partición que se introducen para programas lógicos).

Del mismo modo, la noción de intensionalidad (y la de refuerzo) se extienden para cualquier lenguaje descriptivo, basándose en una definición nueva y formal de

subprograma, aunque el resultado es una formalización bastante incómoda y poco práctica.

Finalmente, la última aproximación se basa en la noción de descripciones proyectables. Definamos primero qué es una descripción proyectable, es decir, una descripción que puede predecir la evidencia futura.

Definición 3.31 (*en la disertación 6.93*) Descripción *k*-Proyectable

Una descripción *k*-proyectable para una cadena x es un programa p sobre un mecanismo descripcional ϕ tal que:

$$\phi(p) = y, \text{ y } \exists w\ l(w) = k : y = xw \text{ (i.e. } x = y_{0..l(x)})$$

w es por tanto la *predicción* de p.

Según el principio MDL, dada cualquier secuencia x, el modelo óptimo en ϕ para ella es x^*. Si x^* es proyectable, es decir, permite predecir los símbolos siguientes de la secuencia x, entonces $\phi(x^*)_{n+1}$ sería la predicción más plausible según la navaja de Occam. Sin embargo, si x^* no es proyectable, no se puede hacer dicha predicción. Por tanto, deberíamos definir un principio MDL modificado. Sólo necesitamos para ello definir una variante proyectable de la complejidad Kolmogorov.

Definición 3.32 (*en la disertación 6.94*) Complejidad Kolmogorov *k*-Proyectable

La *Complejidad Kolmogorov k-Proyectable* de un objeto x dado y sobre un mecanismo descripcional β se define como:

$$K'_\beta(x|y) = \min \{ l_\beta(p) : \exists w\ l(w) = k \text{ tal que } \phi_\beta(<p, y>) = xw) \}$$

donde p denota cualquier programa en β libre de prefijos, y $\phi_\beta(<p, y>)$ denota el resultado de ejecutar p utilizando la entrada y.

Otra cuestión es la extensión proyectable de la complejidad *Kt* de Levin. Para extender la función *LT*, debemos medir Cost(p) de una manera asintótica. Consideremos una máquina ϕ tal que la cinta de salida no puede rectificarse una vez escrita. Cost(p)[..n] se define como el tiempo o pasos de máquina en el que los primeros n símbolos de la salida definitiva están situados al principio de la cinta de salida. Utilizaremos también la siguiente notación: Cost(p)[n..m] = Cost(p)[..m] − Cost(p)[..n]. A partir de aquí se define $LT_\beta(p_x)[n..m] = l(p_x) + log\ \text{Cost}(p_x)[n..m]$ y $LT_\beta(p_x)[..n] = l(p_x) + log\ \text{Cost}(p_x)[..n]$, con lo que se puede ya introducir la variante para LT:

Definición 3.33 (*en la disertación 6.95*) Complejidad Longitud-Tiempo *k*-Proyectable

La *Complejidad Longitud-Tiempo k-Proyectable* de un objeto x dado y sobre un mecanismo descripcional β se define como:

$$Kt'_\beta(x|y) = \min \{ LT_\beta(<p,y>)[..l(x)]-l(y) : \exists w\ l(w) = k \text{ tal que } \phi_\beta(<p, y>) = xw) \}$$

Ya que $LT(<p, y>)$ considera la longitud de y (el conocimiento previo que ya se tiene), esta medida ha debido corregirse con el término $-l(y)$.

Estas variantes parecen, a primera vista, distinguir entre las descripciones que tienen patrón de aquellas que son extensionales. Esto no es así, y la razón es la misma por la cual otros intentos de formalizar la distinción entre patrón y datos han fracasado. Por ejemplo, Bennett introdujo [Bennett 1988] la noción de profundidad lógica con el fin de representar la estructura o complejidad real de un objeto, motivado por el hecho de que la complejidad de Kolmogorov da valores altos para cadenas aleatorias, que no tienen patrón y son estructuralmente sencillas. La profundidad lógica mide la cantidad de tiempo que requeriría una cadena en ser generada de su descripción mínima. Pero, precisamente, Koppel demostró [Koppel 1987] que un concepto llamado "sofisticación" y la profundidad lógica eran equivalente hasta una constante. En la disertación, sin embargo, se muestra que la noción de sofisticación no es válida para distinguir entre patrón y datos, porque parte del patrón se puede transferir a los datos y dejar un intérprete lo más básico posible para estos datos (que serían en realidad los programas). Por tanto, la profundidad lógica tampoco es válida.

Consecuentemente, necesitamos una aproximación diferente para distinguir si cualquier descripción tiene excepciones (parcial o totalmente extensional) o si está compuesta exclusivamente de *patrón* (si todo es estructura o es totalmente intensional). La idea reside en comparar la parte que se usa para todos los datos (en el límite), que es la estructura, con la parte que sólo se usa para una porción de los datos (la excepción).

Definición 3.34 (*en la disertación 6.96*) Equivalencia en el Límite

Una descripción p' es (n,k)-equivalente en el límite a una descripción p sii

$$\exists n \in \mathrm{N},\ n > 0 \text{ y } \exists k \in \mathrm{Z} \text{ tal que } \phi(p')_{n+k..} = \phi(p)_{n..}$$

Informalmente, dos descripciones son equivalentes en el límite si hay un punto a partir del cual sus predicciones siempre coinciden. De aquí podemos dar ya una definición de descripción completamente proyectable.

Definición 3.35 (*en la disertación 6.97*) Descripción Completamente Proyectable

Una descripción p es una descripción completamente proyectable para x dado y sii $<p,y>$ es una descripción ∞-proyectable de x y $\neg\exists p'$ tal que

1. $<p',y>$ es (n,k)-equivalente en el límite a $<p,y>$,
2. $<p',y>$ no es extensionalmente equivalente a $<p,y>$ y,
3. $LT(<p',y>)[n+k..n+k+l(x)] < LT(<p,y>)[n..n+l(x)]$.

La segunda condición de que p' no sea equivalente extensionalmente a p evita que dadas dos o más descripciones totalmente equivalentes, sólo la más corta sería la proyectable. La tercera condición mide que p' es más simple que p.

Esta aproximación final permite formalizar la idea de comprensión, y ayuda a diferenciar entre inducción descriptiva e inducción explicativa, requiriendo ésta última que todas las observaciones sean 'consiliadas' por la teoría, evitando excepciones o casos extensionales.

Se puede definir una variante explicativa de la complejidad de Kolmogorov para dar una variante explicativa del principio MDL, ya sea mediante la función $\Delta(p_x)$ o basándose en las descripciones proyectables. Veamos esta última opción:

Definición 3.36 (*en la disertación 6.100*) Complejidad Explicativa (Versión Proyectable)

La *Complejidad Explicativa* de un objeto x dado y en un mecanismo descripcional β se define como:

$$Et_\beta(x|y) = \min \{ LT_\beta(<p,y>)[..l(x)] - l(y) \text{ tal que } <p,y> \text{ es completamente proyectable } \}$$

Denominaremos $SED(x|y)$ a la Descripción Explicativa más corta para x dado y, es decir, la primera descripción completamente proyectable para x dado y. Lógicamente, $l(SED(x|y)) = Et(x|y)$.

Sin embargo, todavía tenemos que para muchas cadenas, $SED(x)$ será simplemente la descripción "de carrerilla" "repite x indefinidamente" que no supone ninguna comprensión. Una primera idea para evitar este fenómeno sería forzar que la descripción fuera más corta que los datos y decir que los datos no tienen explicación (no pueden ser comprendidos) si no existe dicha descripción. Sin embargo, la mayoría de los datos diarios no son comprimibles y aún así se comprenden.

Otra aproximación es excluir aquellas descripciones generadas por aprendizaje de carrerilla, es decir, la repetición extensional de parte o todos los datos. Esta idea no es nueva y, de hecho, algunos criterios de evaluación como el refuerzo o la validación cruzada están basados en ello. En general, podemos utilizar esta técnica del refuerzo por la derecha:

Definición 3.37 (*en la disertación 6.101*) Estabilidad por la Derecha

Una cadena x es *m-estable por la derecha* en el sistema descripcional β si

$$\forall d, 1 \le d \le m : \text{SED}_\beta(x_{-d}) \text{ es equivalente extensionalmente } \text{SED}_\beta(x)$$

En otras palabras, una cadena x es *m-estable por la derecha* si quitándole hasta m elementos por la derecha, todavía da la misma mejor explicación. Estos m elementos, si se dieran a posteriori, se considerarían refuerzo o confirmación de la explicación, y, si se dan a priori, se consideran redundancia o *pistas* para encontrar la explicación.

Consecuentemente, aunque el aprendizaje de carrerilla puede utilizarse para hacer una descripción extensional completamente proyectable, tanto el refuerzo como la validación cruzada son criterios que evitan fácilmente este fenómeno.

Existe todavía otra razón para apoyar la noción anterior de comprensión/intensionalidad como un principio ontológico. En otras palabras, ¿por

qué debemos evitar el aprendizaje de carrerilla? ¿por qué se debe anticipar? ¿por qué los niños encuentran patrones complejos? [Marcus et al. 1999] ¿por qué estamos genéticamente programados para intentar abrir todas las cajas negras que se nos presentan? Esta búsqueda por hipótesis informativas en vez de las más fáciles y explícitas podría llevar a la fantasía, pero no es peligrosa siempre y cuando el sistema pueda interactuar con el mundo para refutar las hipótesis fantásticas.

Esta informatividad o inversión en las hipótesis fue defendida por Popper para el método científico, y, como hemos visto, es también aplicable a la cognición. Incluso si hiciéramos la suposición de la navaja de Occam, es decir, las cosas en la naturaleza no son complejas innecesariamente, el razonamiento anterior se justifica por el hecho de que, así como cualquier cadena incomprimible suficientemente larga tiene subcadenas comprimibles, la mayoría de las cadenas comprimibles tienen subcadenas incomprimibles, porque cuanto más corto menos útil es comprimir. Si la evidencia se presenta incrementalmente, es mejor invertir en hipótesis más informativas o generales que buscar la óptima para cada trozo, que finalmente no resultará ser parte de la descripción total de la evidencia completa. Este razonamiento se justifica más aún por el siguiente teorema:

Teorema 3.14 (*en la disertación 6.28*) Anticipación

Para cualquier mecanismo descripcional β, existe una constante c que depende exclusivamente de β tal que para toda cadena x de longitud n con $\mathrm{SED}(x) = x^*$ y $l(x^*) = m$ tal que $m < n$, entonces cualquier partición $x = yz$, $l(y) < m - c$ tal que $\mathrm{SED}(y)$ no es equivalente en el límite con x^*.

Como última extensión de las nociones anteriores, hablemos del problema de la incuestionabilidad o certeza de una hipótesis, la cual no depende sólo del criterio de evaluación. Se ha sostenido frecuentemente en la filosofía de la Ciencia y de la inducción que la plausibilidad e incuestionabilidad de una teoría o explicación no sólo depende de las características intrínsecas de la explicación si no de la capacidad de encontrar explicaciones alternativas.

Hagamos formal esta idea. A primera vista parece ser que la estabilidad ya evita esto, pero aún si nos restringimos a descripciones estables, todavía podemos modificar cualquier explicación p con el añadido "Ejecuta p pero imprime un '1' cada cien símbolos" la cual sería comprehensiva para los datos pero diferiría de p en el límite, y sería sólo un poco más larga.

Por esta razón, se extiende la noción anterior de estabilidad y se aplica a descripciones:

Definición 3.38 (*en la disertación 6.102*) Plausibilidad por la Derecha

Una descripción completamente proyectable p para una cadena x es (c,m)-*plausible por la derecha* en el sistema descriptivo β sii

$$\forall d,\ 0 \le d \le m : LT_\beta(\mathrm{SED}_\beta(x_{..d}))[..l(x_{..d})] + c > LT_\beta(p)[..l(x_{..d})].$$

Intuitivamente, una descripción es plausible si es una de las c-mejores explicaciones para x y esto se cumple incluso si quitamos hasta m elementos por la derecha de x.

Una vez extendida la noción de estabilidad, se puede abordar la incuestionabilidad de la siguiente manera:

Definición 3.39 (*en la disertación 6.103*) Incuestionabilidad

Una descripción completamente proyectable p para x es (*c,m*)-*incuestionable* en el sistema descriptivo β sii es (*c,m*)-*plausible* y no existe otra descripción p' (*c,m*)-*plausible* para x.

Esta condición es más restrictiva a medida que c y m aumentan.

Por último, se establece la conexión entre la idea de ganancia de información y la noción de intensionalidad de una descripción, que, como era esperar, se muestra no sólo que las definiciones extensionales tienen ganancia cero sino que las definiciones intensionales, entendidas éstas como definiciones sin excepciones, tienen gran probabilidad de tener ganancia de información alta. En concreto:

Teorema 3.15 (*en la disertación 6.29*)

Dada una descripción eficiente x para unos datos lo suficientemente grandes y, tal que x contiene una *cita* (parte extensional) secuencial Q de una secuencia aleatoria q de y de un tamaño suficiente, en concreto, $l(q)=e > \log^2 l(y)$, entonces x no es intensional y $G(x|y) < 1-e/l(x)$.

Por ejemplo, 1.000 bits de datos con una descripción de longitud 200 bits que contiene una parte extensional de 120 bits no es intensional y $G(x|y) < 0.4$.

Finalmente, se examina rápidamente la relación entre la noción de intensionalidad como se ha presentado en este capítulo y el sentido original de la palabra intensión, más relacionada en la literatura con aspectos de filosofía del lenguaje y el significado.

3.2 Aplicaciones

Los capítulos 3 al 6 han ido introduciendo diferentes conceptos e ideas acerca de la informatividad de teorías, su refuerzo y su intensionalidad. Muchas de las aplicaciones (sobre todo de tipo teórico o explicativo) han ido presentándose a medida que los conceptos se iban definiendo. Entre ellas, se ha visto la diferencia entre implícito y explícito, algunos apuntes sobre el problema de la creatividad y el reconocimiento de qué es descubrir algo, una noción de aprendizaje auténtico, la evaluación de sistemas deductivos, especialmente teorías lógicas de 1er orden, una medida detallada que permite estudiar el crecimiento y revisión de conocimiento y, finalmente, una clarificación de las nociones de explicación, comprensión e incuestionabilidad basadas en la noción de intensionalidad. Aún así, sería extraño que un trabajo acerca de aspectos tan fundamentales de los procesos de inferencia no tuviera todavía más aplicaciones y, deseablemente, de carácter más práctico.

Así, los capítulos 7, 8 y 9 presentan una serie de instrumentalizaciones y aplicaciones de las nociones y construcciones más relevantes de este trabajo.

3.2.1 Evaluación y Generación de Teorías Lógicas

En el artículo "*Complexity-based Induction*" [Conklin and Witten 1994], se presenta una comparación poco imparcial (en mi opinión) de criterios de evaluación, concretamente, el principio MDL basado en la complejidad del modelo y el principio MDL basado en la complejidad de la demostración. Contrarrestaremos en cierta medida las conclusiones de ese artículo contemplando muchos más criterios de evaluación y los mismos ejemplos, considerando muchas más teorías para ellos, teorías que no fueron consideradas por Conklin y Witten porque sus conclusiones habrían sido mucho menos convincentes.

En lo que sigue, se ilustrará la aplicación de las medidas introducidas por Conklin y Witten, algunas otras medidas no consideradas por ellos (pero claramente mejores) y las diferentes medidas que se han presentado en este trabajo.

En concreto, estas medidas son: el grado de generalidad de una teoría, la longitud de la teoría (o MDL de cobertura), el principio MDL descripcional basado en complejidad del modelo, el principio MDL descripcional basado en la complejidad de la demostración, la teoría del refuerzo, intensionalidad y la ganancia de información. La diferencia clásica que haremos entre MDL de cobertura y MDL descripcional es la misma que la presentada por Conklin y Witten, importante por que clarifica cuándo la teoría cubre más hechos que los que se dan en la evidencia (es decir, la teoría generaliza la evidencia) y cuándo simplemente describe los hechos. Por tanto, difieren los resultados del MDL de cobertura que da la teoría más corta que cubre la evidencia y el MDL descripcional que da la teoría que describe más cortamente la evidencia.

Introduzcamos en primer lugar una medida clásica de generalidad para programas lógicos:

Definición 3.40 (*en la disertación 7.104*) El grado de generalización de un programa lógico P con respecto a un conjunto de literales básicos E, denotados $GD(P|E)$, es:

$$GD(P|E) = \text{card } M^+(P) \,/\, \text{card}(E^+)$$

siendo $M^+(P)$ el modelo de P. Si $GD(P|E) < 1$, entonces el programa no cubre todos los ejemplos. Si $GD(P|E) > 1$, que es el caso general, y hasta un cierto límite deseable, el programa generaliza la evidencia. La idea es ajustar $GD(P|E)$ = card(*Total de Ejemplos Positivos Posibles*) / card(Ejemplos Positivos Presentados) pero, obviamente, el total de ejemplos positivos posibles no se sabe a priori.

Pasemos ahora a definir las variantes del principio MDL. En primer lugar debemos computar la longitud de los programas (o complejidad de modelo) de una

evidencia respecto a un programa. Codificaremos (y obtendremos la longitud de) las reglas de un programa en la siguiente manera [Conklin and Witten 1994]:

$l(P) = 1 + \log(v + 1) + 2$ bits por literal + la talla de cada literal

computando la talla de cada literal como $size(l) = a \log(v + c)$, siendo a la aridad del predicado del literal, c el número de constantes en el programa y siendo v el número de variables de la regla con más variables diferentes. El uso de esta medida directamente da lugar al MDL de cobertura.

El principio MDL descripcional para los programas lógicos se define en términos de la complejidad del modelo de la siguiente manera:

$$MDL_1(T \mid E) = L(T) + \log\binom{l(T)}{l(E)}$$

La medida de complejidad de demostración $PC(E|T)$ [Muggleton et al. 1988] [Muggleton et al. 1992] se introduce en el capítulo 4, de la siguiente manera:

Definición 3.41 (*en la disertación 4.28*) **Complejidad de Demostración (PC)**

[Muggleton et al. 1988, 1992]

Dada una teoría T, supongamos que :- G_1, …, G_n es el objetivo actual y la raíz de una rama de éxito del árbol SLD. En este momento se puede seleccionar k reglas donde G_1 (suponiendo reglas de computación por la izquierda) unifica con la cabeza de la regla. Así, se requerirán $\log k$ bits para seleccionar la regla, y se requerirán $\log (c+v)$ bits para las sustituciones, suponiendo programas libres de funciones[8], pero sólo para cada variable de una regla no generativa, es decir, una regla donde la cabeza contiene uno o más variables que no ocurren en el cuerpo de la regla. Llamemos $L_{PC}(a|T)$ la información que se requiere de esta manera para codificar la demostración de un átomo. Por tanto,

$$PC(E|T) = \sum_{a \in E} (L_{PC}(a|T) + 1)$$

A partir de aquí se puede dar la variante del principio MDL de complejidad de demostración:

$$\mathrm{MDL}_2(T|E) = L(T) + PC(E|T)$$

También en el capítulo 4 se estimó $G(E|T)$ para la deducción en teorías lógicas. En este capítulo 7 se necesita una estimación para $G(T|E)$, representando la explicitez de una teoría con respecto a la evidencia. Aunque depende del método de inducción usado, se introduce una aproximación para $G(T \mid <E,B>)$ más o menos apropiada.

Finalmente, la teoría del refuerzo se va a modificar para tener en cuenta la generalidad y el grado de ejemplos positivos y negativos, de la siguiente manera:

[8] [Conklin and Witten 1994] tampoco tienen en cuenta la sustitución entre variables.

Seleccionaremos la variante $m\chi^0$ que se vio en el capítulo 5, porque ponderaba independientemente los valores de ρ^+ y ρ^-. El curso medio que vamos a utilizar lo denominaremos *especializado* y es el siguiente:

$$m\chi' = m\chi^0 \cdot (1 - 0.5f + f \cdot 2^{-GD})$$

donde $f = (n^+ - n^-) / (n^+ + n^-)$, siendo n^+ el número de ejemplos positivos y n^- el número de ejemplos negativos.

Es importante tener en cuenta que esta fórmula puede dar un valor de $m\chi'$ ligeramente mayor que 1. Si $f > 0$ (más ejemplos positivos que negativos) entonces la generalidad se penaliza porque es lo más fácil. Por el contrario, si $f < 0$ (más ejemplos negativos que positivos) entonces la generalidad se favorece porque es más difícil. Lógicamente, si GD = 1 entonces $m\chi' = m\chi^0$.

EJEMPLO:

Vamos a elegir uno de los ejemplos más clásicos en la literatura de ILP, también revisitado en [Conklin and Witten 1994], que aparece originalmente en [Quinlan 1990] y describe la relación de conexión o "alcanzabilidad" en una red. La signatura contiene dos predicados binarios *reach* y *linked*, además de $c = 9$ constantes $\{0,\ldots,8\}$. La teoría del conocimiento previo B se compone de 10 hechos extensionales:

B = { linked(0,1), linked(0,3), linked(1,2), linked(3,2), linked(3,4), linked(4,5), linked(4,6), linked(6,8), linked(7,6), linked(7,8) }

Este conocimiento previo se representa en la Figura 7.1 de la disertación.

CASO 1: Evidencia Completa: todos los ejemplos positivos.

Este es el caso más sencillo, porque podemos aplicar la suposición de mundo cerrado: tenemos todos los ejemplos positivos y el resto es negativo.

En este caso, la evidencia E es una especificación *completa* del predicado *reach* compuesto de 19 hechos sobre 72 combinaciones posibles:

E={ reach(0,1). reach(0,2). reach(0,3). reach(0,4). reach(0,5). reach(0,6). reach(0,8). reach(1,2). reach(3,2). reach(3,4). reach(3,5). reach(3,6). reach(3,8). reach(4,5). reach(4,6). reach(4,8). reach(6,8). reach(7,6). reach(7,8) }

Vamos a contemplar las teorías que se muestran en la tabla 1:

Teoría	Programa	Comentario
T_1	reach(X,Y)	$T_1 = \top$
T_2	reach(0,1). reach(0,2). reach(0,3). reach(0,4). reach(0,5). reach(0,6). reach(0,8). reach(1,2). reach(3,2). reach(3,4). reach(3,5). reach(3,6). reach(3,8). reach(4,5). reach(4,6). reach(4,8). Reach(6,8). Reach(7,6). reach(7,8)	$T_2 = \perp + E$
T'_2	reach(0,X). Reach(3,X). reach(X,8). reach(1,2). reach(4,5). Reach(4,6). reach(7,6).	T'_2 = generalización simple cuando hay más de 5 hechos.
T_3	reach(X,Y) :- linked(X,Y). reach(0,2). reach(0,4). reach(0,5). reach(0,6). reach(0,8). reach(3,5). reach(3,6). reach(3,8). reach(4,8).	
T_4	reach(X,Y) :- linked(X,Y). reach(X,Y) :- linked(X,Z). (T'_4)	La segunda cláusula subsume a la primera.
T_5	reach(X,Y) :- linked(X,Y). reach(X,Y) :- linked(X,Z), linked(Z,Y). reach(0,5). reach(0,6). reach(0,8). reach(3,8).	
T_6	reach(X,Y) :- linked(X,Y). reach(X,Y) :- linked(X,Z), reach (Z,Y).	La teoría 'intuitiva'.

Tabla 1. Teorías para la relación de 'alcanzabilidad'.

Las teorías T'_2 y T'_4 no fueron contempladas por [Conklin and Witten 1994]. Veamos el resultado de las medidas presentadas a estas teorías en la tabla 2:

T	L(T)	GD	Consilte. (sin exceps.)	*Ganancia*	Curso Medio ($m\chi$)	Spec. ($m'\chi$)	L(E|T)	**MDL_1**	PC(E|T)	**MDL_2**
T_1	11.5	3.8	Sí	0.57	≈ 1	0.57	56.7	68.2	120.5	**132.0**
T_2	159.5	1	No	0.02	= 0.5	0.5	0	159.5	80.7	240.2
T'_2	60.3	1.52	No	0.59	0.88	0.75	24.3	84.6	100.9	161.2
T_3	111.7	1	No	0.09	0.76	0.76	0	111.7	96.3	208.0
T_4	43.7	2,53	No	0.58	≈ 1	0.67	43.4	87.1	110.6	154.3
T'_4	23.3	2,53	Sí	0.75	≈ 1	0.67	43.4	66.7	123.3	133.9
T_5	94.5	1	No	0.39	0.886	0.89	0	94.5	101.9	196.5
T_6	53.8	1	Sí	0.68	0.999	**0.999**	0	**53.8**	106.1	160.0

Tabla 2. Valores para los diferentes criterios estudiados para evidencia positiva total.

El resultado de la tabla es auto-explicativo. Aunque MDL_1 y $m'\chi$ eligen T_6 como la mejor teoría, es el refuerzo $m'\chi$ quien la muestra con mayor claridad.

Además, la aproximación a la ganancia también proporciona información útil acerca de las teorías consideradas. Atendiendo al criterio de olvido visto en los capítulos 4 y 5 como el producto de la ganancia y el criterio de plausibilidad podemos seleccionar las mejores hipótesis según el esfuerzo y la plausibilidad. Por ejemplo, si en el caso anterior sólo pudiéramos retener 3 teorías, elegiríamos (usando $m'\chi$ como

un criterio de plausibilidad) las tres teorías que se muestran en negrita en la siguiente tabla:

T	Ganancia	m'_χ	OC
T_1	0.57	0.57	0.32
T_2	0.02	0.5	0.01
T'_2	0.59	0.75	**0.44**
T_3	0.09	0.76	0.07
T_4	0.58	0.67	0.38
T'_4	0.75	0.67	**0.50**
T_5	0.39	0.89	0.35
T_6	0.68	0.999	**0.68**

CASO 2: Evidencia Parcial: Muestra Positiva Parcial

Este el caso más habitual cuando se aprende a partir de evidencia positiva: sólo una parte de ella aparece. Para estudiar este caso, el ejemplo anterior se modifica a una evidencia E diferente con 12 ejemplos en vez de los 19 totales y se mantiene el conocimiento previo B.

E = { reach(0,3). reach(0,4). reach(0,5). reach(0,8). reach(3,2). reach(3,4). reach(3,5). reach(3,8). reach(4,6). reach(4,8). reach(6,8). reach(7,8) }

Las teorías a considerar son las mismas que las anteriores exceptuando T_2, T_3 y T_5:

Teoría	*Programa*
T_1	reach(X,Y)
T_2	reach(0,3). reach(0,4). reach(0,5). reach(0,8). Reach(3,2). reach(3,4). reach(3,5). reach(3,8). reach(4,6). reach(4,8). Reach(6,8). reach(7,8).
T'_2	reach(0,X). reach(3,X). reach(X,8). reach(4,6).
T_3	reach(X,Y) :- linked(X,Y). reach(0,4). reach(0,5). reach(0,8). reach(3,5). reach(3,8). reach(4,8).
T_4	reach(X,Y) :- linked(X,Y). reach(X,Y) :- linked(X,Z).
T_5	reach(X,Y) :- linked(X,Y). reach(X,Y) :- linked(X,Z), linked(Z,Y). reach(0,5). reach(0,8). reach(3,8).
T_6	reach(X,Y) :- linked(X,Y). reach(X,Y) :- linked(X,Z), reach (Z,Y).

Tabla 3. *Nuevas teorías para la relación de 'alcanzabilidad' y su correspondiente* m_χ.

La tabla de resultados varía ahora significativamente:

T	L(T)	GD	Consilte (sin exceps.)	Ganancia	Curso Medio ($m\chi$)	Spec. ($m'\chi$)	L(E|T)	**MDL_1**	PC(E|T)	**MDL_2**
T_1	**11.5**	6	Sí	0.57	≈ 1	0.52	43.8	**55.3**	76.1	**87.6**
T_2	101.1	1	No	0.02	= 0.5	0.5	0	101.1	43.0	144.1
T'_2	35.4	2.17	No	≈ 1	0.91	0.66	23.2	58.6	58.9	94.3
T_3	81.9	1.33	No	0.13	0.74	0.66	10.8	92.7	94.1	176.0
T_4	43.7	4	No	0.58	≈ 1	0.56	36.0	79.7	70.9	114.6
T'_4	23.3	4	Sí	0.75	≈ 1	0.56	36.0	59.3	77.9	101.2
T_5	84.5	1.25	No	0.43	0.836	0.77	8.83	93.3	70.3	154.8
T_6	53.8	1.58	Sí	0.68	0.987	**0.82**	15.6	69.4	81.9	135.7

Tabla 4. *Valores para los diferentes criterios estudiados para evidencia positiva parcial.*

Se puede observar que cuando la evidencia se reduce, ambas variantes del principio MDL descartan la teoría T_6 y dan mejores valores para otras teorías. Por el contrario, el refuerzo todavía la selecciona como la mejor teoría.

CASO 3: Evidencia Parcial: Evidencia Positiva y Negativa

Finalmente, consideremos evidencia negativa. Ya que el aprendizaje de evidencia positiva y negativa es más fácil que de evidencia positiva sólo, en teoría se deberían obtener resultados más claros que en los casos anteriores.

Tenemos la misma evidencia positiva que en el caso anterior y evidencia negativa:

E^- = { reach(8,3). reach(5,4). reach(0,7). }

Consideraremos las mismas teorías, por lo que la ganancia y las longitudes se mantienen. El refuerzo positivo también se mantiene y no hay refuerzo negativo para T_2, T_3, T_5 y T_6. Sin embargo, para el resto de teorías, los resultados cambian, como se muestra en la siguiente tabla:

T	L(T)	GD	Consilte. (sin exceps.)	Ganancia	Curso Medio ($m\chi^0$)	Spec. ($m'\chi^0$)	L(E|T)	**MDL_1**	PC(E|T)	**MDL_2**
T_1	11.5	6	Sí	0.57	0.78	0.50	43.8	**55.3**	76.1	**87.6**
T_2	101.1	1	No	0.02	= 0.5	0.5	0	101.1	43.0	144.1
T'_2	35.4	2.17	No	≈ 1	0.87	0.79	23.2	58.6	58.9	94.3
T_3	81.9	1.33	No	0.13	0.74	0.68	10.8	92.7	94.1	176.0
T_4	43.7	4	No	0.58	0.94	0.63	36.0	79.7	70.9	114.6
T'_4	23.3	4	Sí	0.75	0.94	0.63	36.0	59.3	77.9	101.2
T_5	84.5	1.25	No	0.43	0.836	0.79	8.83	93.3	70.3	154.8
T_6	**53.8**	1.58	Sí	0.68	0.987	**0.86**	15.6	69.4	81.9	135.7

Tabla 5. *Valores para los diferentes criterios para evidencia positiva y negativa parcial.*

Nótese que tanto MDL_1 y MDL_2 no cambian porque no consideran los errores. Esto se debe a que $L(E|T)$ debe decir qué hechos de $M(T)$ están realmente en E^+. Ya que el objetivo del principio es descripcional, no es relevante si parte de la información se requiere para decir que un ejemplo todavía no aparecido (e en $M(T)$ pero e no está en E^+ y tampoco en E^-) o si la información se requiere para los casos donde (e en $M(T)$ y e en E^-).

Un partidario del principio MDL podría decir que T_6 es la teoría más corta sin errores. Esto es cierto en este caso pero haría el principio MDL inútil precisamente en los casos donde ha tenido más éxito, el aprendizaje de datos ruidosos. Una mejor idea es rectificar el principio de MDL con la proporción de errores, dando un nuevo principio MDL:

$$MDL^{+,-}_1 = MDL_1 \cdot 2^{\alpha \cdot e(T)}$$

siendo $e(T)$ el ratio de error (ejemplos negativos cubiertos / ejemplos positivos cubiertos). El resultado con distintos factores se muestra en la siguiente tabla:

T	$e(T)$	MDL_1	$MDL^{+,-}_1$ ($\alpha = 1$)	$MDL^{+,-}_1$ ($\alpha = 5$)
T_1	0.25	**55.3**	65.8	131.5
T_2	0	101.1	101.1	101.1
T'_2	0.083	58.6	62.1	104.4
T_3	0	92.7	92.7	92.7
T_4	0.083	79.7	84.4	106.4
T'_4	0.083	59.3	**62,8**	79.2
T_5	0	93.3	93.3	93.3
T_6	0	69.4	69.4	**69.4**

Claramente, a medida que el valor de α es mayor, la nueva medida es menos robusta a los errores.

Finalmente, consideremos el caso contrario, 3 ejemplos positivos y 12 negativos ($f = -0.75$). En este caso, la teoría extensional T_2 con refuerzo medio = 0.5 y GD = 1 y la misma teoría T_6 se tiene:

reach(X,Y) :- linked(X,Y). : $\rho = 1 - 2^{-3} = 0.875$

reach(X,Y) :- linked(X,Z), reach (Z,Y). : $\rho = 1 - 2^{-2} = 0.75$

un curso medio $m\chi = (0.875 \cdot 2 + 0.875 \cdot 0.75 \cdot 1) / 3 = 0.8$.

Con una $GD = 19 / 3 = 6.33$ tenemos:

$$m'\chi(E|T_2) = m\chi(E|T_2) \cdot (1 + 0.5 \cdot 0.75 - 0.75 \cdot 2^{-GD}) = 0.5$$

$$m'\chi(E|T_6) = m\chi(E|T_6) \cdot (1 + 0.5 \cdot 0.75 - 0.75 \cdot 2^{-GD}) = 1.09$$

Es razonable que $m'\chi(T_6)$ se incremente debido al hecho de que una teoría general haya sobrevivido a una evidencia mayoritariamente negativa.

CASO 4: Evidencia Ruidosa

El principio MDL, como se ha comentado, se ha aplicado satisfactoriamente a datos ruidosos. Ya que el objetivo es comprimir, se pueden añadir algunos parches extensionales a la teoría mientras que el ratio de compresión global se mantiene alto. Sin embargo, el principio MDL es ciego al grado de errores de la evidencia. La solución habitual en un mecanismo descripcional universal es anotar las excepciones aparte, sean positivas o negativas, y, por tanto, no hay excepciones intrínsecas, y se penalizan por el mismo incremento de tamaño en la teoría. Pero en el caso de teorías lógicas de Horn, no podemos quitar una consecuencia para parchear una teoría, es decir, no podemos expresar $M(T) - f$. Esto representa una forma no uniforme de considerar la evidencia positiva excluida (simplemente parchéese) y la evidencia negativa incluida. La solución puede ser una ponderación como $MDL^{+,-}_{1}$.

No obstante, el principio MDL es ciego en otro sentido. Da un simple valor para la teoría, y no se puede saber finalmente qué porcentaje de los datos está cubierto extensionalmente, ya que es difícil saber si hay muchos errores o no en la teoría.

Por el contrario, el refuerzo y la intensionalidad sirven para distinguir el ratio de error de una teoría, y compararlo con el ratio de error esperado para la evidencia. Este es precisamente el resultado más práctico de la *complejidad explicativa* y del principio SED introducidos en el capítulo 6. Dados ciertos datos x, si tenemos una expectativa de ruido de cerca del 3%, debemos buscar sólo descripciones cuya parte extensional es $\Delta(p_x) \approx l(x) \cdot 0.03$. Es importante darse cuenta de que el principio MDL da un ratio de excepciones *incontrolable* e *impredecible*, que sólo depende de los datos.

Conclusiones

En el marco del aprendizaje incremental, un criterio intensional es menos conservativo que el principio MDL, y consecuentemente minimiza el número total de 'cambios mentales' (aunque estos cambios son más radicales) cuando los datos son perfectos. Se podría decir que el principio MDL se corresponde con la filosofía de Kuhn de los paradigmas cambiantes; cuando el número de excepciones es muy grande, el paradigma debe cambiarse. Por contra, un criterio intensional anticipa esta necesidad ya que cualquier excepción fuerza la revisión del modelo. En resumen, es más anticipativo en el sentido del capítulo 4.

La teoría del refuerzo está de alguna manera entre los dos extremos. Este compromiso se ha mostrado como un criterio mucho más fiable que el principio MDL. Aunque la comparación se debería realizar sobre muchos más ejemplos, los resultados que se muestran aquí pueden esperarse en general, debido a las justificaciones teóricas dadas en el capítulo 6.

Para acabar, la sección 3 de este capítulo 7, que no se discutirá aquí en el resumen, muestra como el refuerzo y la intensionalidad pueden combinarse para guiar un algoritmo de aprendizaje automático. En primer lugar, se muestra que un algoritmo de enumeración es compatible con una búsqueda de ganancia, porque los datos sólo

se usan para comprobar las hipótesis. En segundo lugar, una aproximación guiada por los datos se puede construir por medio de la programación genética, siendo el criterio de selección (criterio de olvido) una combinación de la optimalidad del programa (el individuo) con la ganancia (genotipo rico o inusual). Más aún, el carácter aleatorizado de la programación genética permite la generación de hipótesis informativas.

3.2.2 Medición de Capacidades Intelectuales

El capítulo 8, *Measurement of Intellectual Abilities*, presenta la aplicación más llamativa y sugerente de esta tesis.

La Inteligencia Artificial se ha esforzado por imitar el comportamiento humano en muchos aspectos, bajo el eslogan "*La Inteligencia Artificial es aquello que si se hiciera por los seres humanos requeriría inteligencia*", que ha fomentado la visión de que "*la inteligencia humana subsume la inteligencia de las máquinas*" [Bradford and Wollowski 1995] en vez de la visión más abierta y realista de que "*los robots serán más inteligentes de lo que los humanos son*" [Moravec 1998]. Finalmente, el Test de Turing (TT) se ha entendido habitualmente como un test efectivo (y no como un ejercicio filosófico). Esta mala interpretación, juntamente con su celebridad, ha motivado que no haya habido el esfuerzo necesario por diseñar nuevos tests de inteligencia alternativos. El TT ha eclipsado incluso algunas propuestas tan reputadas como los primeros trabajos de Simon sobre la relación entre los tests IQ y IA [Simon and Kotovsky 1963], algunas aproximaciones heurísticas para resolver los problemas de analogía de los tests IQ [Evans 1963] y la sugerencia de Chaitin "*desarróllense definiciones formales de inteligencia y medidas de sus varios componentes*" [Chaitin 1982], basándose en la complejidad descripcional. Incluso la llamada de Johnson "*Se requiere: Un Nuevo Test de Inteligencia*" [Johnson 1992] ha sido respondida con formalizaciones del TT [Bradford and Wollowski 1995] en vez de diseñar nuevas propuestas.

Como cualquier otra disciplina, la inteligencia artificial necesita una medida efectiva de su característica más relevante, una medida gradual y detallada de inteligencia. Como se discute en la disertación, una medida científica de inteligencia debería seguir los siguientes requerimientos: No booleana (gradual), Factorial, No antropomórfica, Computacionalmente fundada y Significativa. Veamos qué problemas técnicos y cómo abordarlos para construir un test que cumpla estos requisitos.

La psicometría se ha esforzado en mostrar que no es absurdo medir la solución 'correcta'. Su respuesta es que la gran mayoría coincide con la solución propuesta *porque no hay soluciones alternativas de complejidad similar*, y, consecuentemente, es la más plausible. Sin embargo, esta afirmación se ha hecho siempre desde un punto de vista subjetivo e informal.

Como se vio en el capítulo 6, se introdujo la noción de estabilidad para evitar las descripciones repetitivas (o de carrerilla). No obstante, existía la necesidad de evitar

patrones extra que pudieran hacer que las predicciones difirieran. Por esta razón, se extendió la noción de estabilidad y se aplicó a descripciones, dando lugar a la noción de plausibilidad por la derecha (Definición 3.38) y a la noción de incuestionabilidad (Definición 3.39).

Esta restricción a descripciones incuestionables permite generar problemas cuya solución no depende de la opinión de un sujeto exterior sino que se deriva matemáticamente. Además, un agente inteligente sólo puede pensar que una descripción es incuestionable cuando tiene la suficiente habilidad de encontrar explicaciones alternativas y, tras un gran esfuerzo, no lo consigue. En este sentido la inteligencia puede verse como el medio más importante de aumentar la plausibilidad y confidencia de las explicaciones, y, consecuentemente, la ontología de un sistema 'inteligente'.

Una vez nos centramos en obtener cadenas tales que su $SED_\beta(x)$ sea (c,m)-incuestionable como ejercicios para el test, necesitamos además discernir la complejidad o la dificultad de cada ejercicio, con el fin de poder dar un conjunto de ejercicios de diferente complejidad. La idea es relacionar esta complejidad con la complejidad explicativa (Et):

Definición 3.42 (*en la disertación 8.111*) **Comprensibilidad (Versión Corregida).** Una cadena x es *k-difícil (*o *k-incomprensible*) dada y, denotado por $incomp(x|y)$, en un sistema descripcional β sii k es el menor número entero positivo tal que:

$$Et_\beta(x|y) \cdot G(\mathrm{SED}(x|y) \mid \langle x,y\rangle) \leq k \cdot \log l(x)$$

Esta ponderación mide la complejidad real de encontrar $\mathrm{SED}(x)$ de x, porque descripciones de la manera "repite x indefinidamente" que tienen Kt alto (para citar a x) se corrigen por G, pero la longitud de x es todavía importante.

Estamos ya preparados para construir un test genérico de habilidad de comprensión mediante la generación de series de cadenas de comprensibilidad creciente. No obstante, como se ha dicho, es importante que la respuesta sea incuestionable, porque si no, la respuesta sería una elección arbitraria del examinador. Una manera sencilla de conseguir esto es dar información redundante para hacer la respuesta incuestionable. Sin embargo, no se puede dar demasiada redundancia, porque si no, los problemas se harían demasiado largos. Por ejemplo, la serie "a, c, c, a, c, c, c, a, c, c, c, c, a, ..." parece continuar por la serie "c, c, c, c, c, a, c, ...", por lo que parece redundante presentar más símbolos de los necesarios.

La medición que se va a presentar a continuación requiere la colaboración del sujeto, que debe emplear todos sus recursos para realizar el test. No es relevante si el sujeto entiende el objetivo del test o simplemente está programado para ello. Es incluso más imparcial si el sujeto desconoce que es un test de inteligencia. Idealmente, el sujeto sólo debe conocer el lenguaje y las cuestiones que componen el test:

Con estas clarificaciones sobre la naturaleza del test, definimos la inteligencia de un sistema *S* cualquiera como el valor que resulta de aplicarle el siguiente test:

Definición 3.43 (*en la disertación 8.112*) **C-Test.** Se selecciona un sistema descriptivo β suficientemente expresivo e imparcial, compuesto de un alfabeto o símbolos Ω_β y un conjunto de operaciones Θ_β para manipular estos símbolos, y su correspondiente *coste* (o longitud). Se proporciona (o programa) a *S* el alfabeto, las operaciones y el coste.

Dependiendo de la inteligencia esperada del sistema se selecciona un rango suficientemente amplio 1..*K* de dificultad. Para cada $k = 1..K$ se eligen aleatoriamente *p* secuencias $x^{k,p}$, siendo *k-incomprensibles*, *c-plausibles*, *c-m-incuestionables* y *d-estables* con $d \geq r$, siendo *r* el número de símbolos redundantes de cada ejercicio.

Se mide la inteligencia del sistema pretendidamente inteligente *S* de la siguiente manera:

Las cuestiones son las secuencia $K \cdot p$ sin sus $d - r$ elementos ($x^{k,p}{}_{\neg(d+r)}$). Se proporcionan a *S* y se pregunta por el siguiente elemento según la mejor explicación que es capaz de construir con Ω_β y Θ_β. Se deja a *S* un tiempo fijo *t* y se registran sus respuestas: $guess(S, x^k{}_{-d+r+1})$.

El resultado de este test de comprensibilidad (o C-test) se mide como:

$$I(S) = \sum_{k=1..K} k^e \cdot \sum_{i=1..p} hit\left[x^{k,i}_{-d+r+1}, guess(S, x^{k,i}_{-d+r+1})\right]$$

la función *hit* se mide habitualmente como $hit(a,b) = 1$ si $a = b$ y 0 en otro caso (los valores negativos pueden usarse para penalizar errores). El valor *e* es simplemente para ponderar las cuestiones difíciles (si se elige $e = 0$ todas las cuestiones tienen el mismo valor).

De una manera informal, "el test mide la habilidad de encontrar la mejor explicación (una descripción completamente proyectable sin descripciones alternativas completamente proyectables de complejidad comparable) para secuencias de comprensibilidad creciente en un tiempo fijo".

Una característica relevante del test es que, aunque se supone que el sujeto es un sistema descriptivo universal particular ϕ_s con un conocimiento previo (experiencia vital) B_s, se le da un sistema descriptivo β sobre él, lo cual minimiza en gran medida la influencia de la diferencia entre las computaciones realizadas por ϕ_s y otro sujeto ϕ_t, es decir, la diferencia entre $Et_s(x|<B_s, \beta>)$ y $Et_t(x|<B_t, \beta>)$. Esto hace posible que las nociones de plausibilidad e incuestionabilidad sean similares para ambos sujetos.

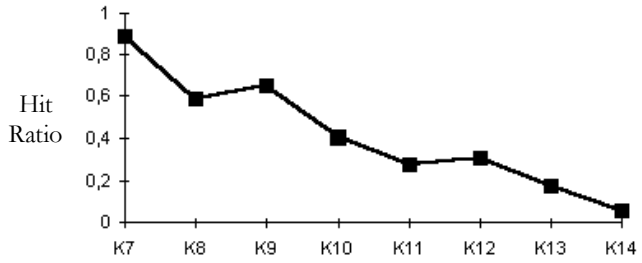
El test anterior es aplicable a cualquier sistema cuyo nivel de inteligencia esté en cuestión. Seleccionando apropiadamente el sistema descriptivo y el resto de parámetros del test, se puede utilizar para seres humanos, animales, computadoras,

seres extraterrestres y cualquier colección de los anteriores trabajando conjuntamente.

Aunque la Definición 3.43 evalúa una sola habilidad, hay todavía muchas maneras de realizar un test específico. En [Hernández-Orallo and Minaya-Collado 1998] el test se implementó utilizando una máquina abstracta similar a una máquina de estados. A partir de aquí, se generaron una variedad de cadenas de diferente *comprensibilidad* en esa máquina. Aunque el conjunto de números *k*-potentes de longitud como máximo *n* puede obtenerse en tiempo polinómico en *n* (véase una demostración en [Li and Vitányi 1997]), el coste de $O(n^k)$ fuerza el uso de algunas heurísticas para ello. De la misma manera, *G* tuvo que ser aproximada. Finalmente, se aplicó una criba para obtener sólo secuencias *c-plausibles*, *c-m-incuestionables* y *d-estable*. Se muestran más detalles en el apéndice del capítulo 8.

El mismo trabajo presenta los resultados de aplicar el test a 65 sujetos de la especie Homo Sapiens Sapiens con edades comprendidas entre 14 y 32 años, juntamente con un test clásico de inteligencia, el *European IQ Test*. La correlación entre ambos tests fue de 0.77. Este valor sólo justifica un estudio más exhaustivo sobre grupos más grandes y con variaciones derivadas de la Definición 3.43. Por el momento, esta evidencia psicométrica es vital para una teoría formal de medición de inteligencia, ya que, atendiendo a Brand [Brand 1996], la correlación con los tests IQ es una condición necesaria (pero no suficiente) para una buena medida de la inteligencia.

Otro resultado experimental importante que se muestra en la figura inferior es que la relación entre el ratio de aciertos y la *k*-incomprensibilidad es lineal, lo que sugiere que la comprensibilidad estima realmente la dificultad de cada cadena:



Lógicamente, es poco interesante saber que el Homo Sapiens medio es capaz de 'entender' secuencias de incomprensibilidad = 10 en un tiempo razonable. De manera similar, no es de esperar (por el momento) que los tests IQ tradicionales contrastados y ampliamente usados se sustituyan por estos C-tests. No obstante, este trabajo podría significar un hito en los fundamentos teóricos de la psicometría porque es la primera medida de un factor de inteligencia que está basado teóricamente y no usando al Homo Sapiens como referencia.

Sin embargo, no es la inteligencia humana sino la no humana lo que es urgente medir. Una declaración formal de lo que se espera de un sistema inteligente debería permitir dos cosas importantes: derivar sistemas más inteligentes a partir de una especificación más concreta y, en segundo lugar, evaluarlos. La Definición 3.43 proporciona un primer paso para estas dos cosas, una escala detallada para medir el progreso (en un factor de inteligencia) de sistemas genéricos en IA. Como cualquier otro campo de la ciencia, un gran avance en una disciplina ocurre cuando uno de sus aspectos fundamentales puede medirse de una manera efectiva y justificada. La Inteligencia Artificial, como ciencia, requiere medidas de inteligencia, o al menos, medidas de sus diferentes factores.

Los sistemas de IA modernos son mucho más funcionales que los sistemas de los sesenta o los setenta. Resuelven problemas de una manera automática que antes requerían la intervención humana. Sin embargo, estos problemas complejos se resuelven porque los diseñadores del sistema encuentran una solución metódica, no porque los sistemas actuales sean más inteligentes que sus predecesores. Nadie se preocupa de cuán funcionales son estos sistemas para otros tipos de problemas, ya que "*es más fácil evaluar sistemas que hacen cosas específicas que evaluar sistemas que hacen tareas más generales*" [Nilsson 1995]. El olvido actual de resolvedores de problemas generales puede ser *tecnológicamente* correcto con una disciplina que demanda a gritos aplicaciones, pero no es justo con el nombre fundacional de la IA. "*Ya es hora de distinguir entre los programas generales e inteligentes y los sistemas de actuación especial*" [Nilsson 1995].

Este objetivo inicial de ser más general está representado hoy en día por dos subcampos de la inteligencia artificial: el razonamiento automático (demostración automática de teoremas, ATP) y el aprendizaje automático. El campo ATP es capaz hoy en día de resolver complejos problemas de diferentes campos de las matemáticas. El gran avance de las últimas décadas se debe fundamentalmente a la existencia de conjuntos de problemas para comparar diferentes sistemas. Estas colecciones de problemas han evolucionado y crecido hasta convertirse en librerías completas y extensas de problemas de demostración de teoremas, como la TPTP [Suttner and Sutcliffe 1996]. El campo del aprendizaje automático está tomando también un carácter más experimental y diferentes sistemas (de diferentes paradigmas) se evalúan atendiendo a los problemas clásicos de la literatura.

Seleccionando un lenguaje representacional apropiado podríamos calcular la complejidad teórica de los problemas que componen estas librerías. Por ejemplo, si se selecciona la lógica de primer orden como mecanismo descriptivo universal, se podría medir la complejidad de la evidencia, la hipótesis (explicación) y conocimiento previo de la misma manera, mediante el uso, por ejemplo, de una medida de la longitud de programas lógicos, como las que se han visto en el capítulo 7. Esto permitiría, por ejemplo, dar la complejidad teórica de los problemas que generalmente se pasan a los sistemas ILP con el fin de ver cuán inteligentes son.

Finalmente, la sección 7 del capítulo 8, la cual no detallaremos aquí, estudia la medición de otros factores, inductivos (aplicabilidad del conocimiento, contextualización, construcción de conocimiento) o deductivos (habilidad de cálculo, habilidad de resolución de problemas, habilidad derivacional) bajo las mismas condiciones con las que se ha diseñado el C-test.

Después de estos resultados y expectativas, el Test de Turing se re-examina en la sección 8 y se reduce a su carácter filosófico e incluso metafórico original. Comparado con el C-test, se reconoce la importancia del TT, pero asimismo se resaltan las agudas deficiencias de su mala interpretación y encarnaciones como el Loebner Prize.

3.2.3 Aplicaciones en Perspectiva

El capítulo 9, *Prospective Applications*, incluye algunas propuestas de aplicación que, en su mayor parte, están todavía en una fase teórica. Se incluyen nuevas herramientas y adaptaciones a campos muy diversos de los conceptos que se han presentado en este trabajo. Por tanto el término 'en perspectiva' indica que no hay una segunda fase de implementación y experimentación de los conceptos y modelos que se presentan, que mostrarían, al fin y al cabo, la validez de los mismos.

Concretamente, en la sección 2 de este capítulo se muestran las aplicaciones de la ganancia de información a los *sistemas de información* (bases de datos). Después de una breve descripción de la minería de datos descriptiva, que es una aplicación de las técnicas de aprendizaje automático para obtener conocimiento de bases de datos, se estudia la posibilidad y utilidad de la minería de datos no predictiva. Para ello, se retoman las medidas de representación óptima vistas en los capítulos 3 y 4 con el fin de discutir cuál sería la representación óptima en bases de datos deductivas, para mejorar el rendimiento de las operaciones de la base de datos dependiendo de qué operaciones son más frecuentes y el grado de regularidad de los datos. Una vez el nivel físico se separa de la cuestión lógica, las relaciones intensionales que se encuentran en una base de datos a un nivel superior son mucho más importantes para la calidad de datos de un sistema, con el fin de controlar la consistencia y redundancia de los datos. Finalmente, se reconoce que tanto los procesos deductivos como los inductivos (y su integración) serán cada vez más importantes en las bases de datos futuras, que serán conocidas como sistemas de conocimiento.

En la sección 3, se reconsideran las características de validación y mantenimiento de sistemas software, bajo la analogía entre la ingeniería del software y la filosofía de la ciencia o, lo que es más preciso, entre la construcción de software y el aprendizaje automático. En particular, las medidas de refuerzo del capítulo 5 se adaptan para definir una medida de la 'predictividad' del software, que se identifica con la validación del software, para representar la estabilidad de un sistema. Una medida inversamente relacionada, la probabilidad de modificación, también se obtiene para cada componente y para el sistema completo. Se discute la aplicación en la práctica

de estas medidas. A partir de aquí, se presentan varios modelos de coste de mantenimiento, basados en una combinación detallada de predictividad y modificabilidad. Así se estudian teóricamente diferentes topologías de disposición del software. Las topologías jerarquizadas, especialmente las que son confluentes en la base tales como los árboles o los retículos implican menos costes de mantenimiento. Más aún, se confirman algunas expectativas intuitivas, como que los sistemas comprimidos y los modelos coherentes (sin parches o excepciones) son claramente más mantenibles.

En la sección 4 se esbozan otras aplicaciones, especialmente relacionas con la interacción y el entendimiento mutuo, algunas cuestiones relacionadas con el significado y el lenguaje, y sus aplicaciones para la comunicación de agentes. Los lenguajes fijos tienen expresividad limitada y sería bastante sorprendente que un lenguaje estándar (como los que se intentan desarrollar) pudieran ser usados por sistemas que se interrelacionan en mundos reales o virtuales. Lenguajes diferentes y lenguajes cambiantes deben ser usados para comunicar con agentes diferentes que tengan inteligencia y conocimiento divergentes, y la comunicación se debe adaptar a estas situaciones. En este contexto, la diferencia entre explícito e implícito, formalizada y clarificada en este trabajo podría tener aplicaciones fructíferas en el área del procesamiento del lenguaje natural. La noción formal de comprensibilidad introducida en el capítulo 6 se ha mostrado también como un aspecto importante a explotar para el entendimiento del lenguaje.

Este último aspecto, la comprensibilidad, podría aplicarse también a los sistemas de conocimiento, porque la inteligibilidad es uno de los factores que afectan la aceptación y la popularización de estos sistemas, ya que muchos de ellos son completamente crípticos para el usuario [Kodratoff 1994].

Las bases de datos y los sistemas software son dos tipos extremos de los sistemas de conocimiento. Los sistemas basados en conocimiento, los sistemas de procesamiento del lenguaje natural y las nuevas tendencias de agentes de información o agentes software están trazando un espectro cada vez más amplio de uso de los métodos de razonamiento inductivos, deductivos, analógicos y abductivos. Más aún, las técnicas anticipativas y perezosas se están empezando a combinar. En este panorama, las aplicaciones de las medidas y otras nociones diferentes presentadas en este trabajo parecen todavía más prometedoras en un futuro cercano.

4. Conclusiones

La determinación de abordar los procesos de inferencia desde un punto de vista no estrictamente semántico parecía arriesgado a primera vista. Hasta qué punto se puede llegar mediante medidas de evaluación realizadas con herramientas descripcionales, computacionales y numéricas era una cuestión al principio de este trabajo. Existían algunas aproximaciones semánticas o lógicas que habían intentado entender infructuosamente diferentes procesos de inferencia de una manera unificada y consistente, bajo algún tipo de marco lógico unificado. Sin embargo, no existían intentos válidos para conocer el resultado de los diferentes procesos de inferencia; en otras palabras, un estudio del qué y del cuánto y no del cómo, que debería ser secundario. En mi opinión, por tanto, una aproximación basada en la evaluación (en nuestro caso computacional y numérica) merecía ser estudiada.

Los resultados han sido, afortunadamente, bastante satisfactorios, atendiendo a las expectativas hechas al principio de esta tesis. Aspectos tan importantes en procesos de inferencia como la explicitez, implicitez, novedad, información intermedia, optimalidad de representación, informatividad, extensionalidad, intensionalidad, plausibilidad y confirmación han sido recogidos por estas medidas. Además, las herramientas han sido exclusivamente definiciones y propiedades básicas de la complejidad descripcional (Kolmogorov) por una parte, y una sencilla teoría numérica del refuerzo por otra. En otras palabras, las medidas que se han presentado (excepto el difícil concepto de intensionalidad) se han mantenido sin entrar en matemáticas intrincadas. La razón se puede encontrar en mi convicción de mantener las cosas tan simples como sea posible y, por supuesto, mi incapacidad declarada de utilizar algunas técnicas matemáticas sofisticadas que podían haber sido útiles en algunas ocasiones. En particular, algunos resultados de la teoría descriptiva de conjuntos, topología, teoría de modelos, y complejas distribuciones de probabilidad deberían aplicarse, de una manera unificada, a diferentes procesos de inferencia. El uso de algunas de ellas habría dado a la tesis otro ámbito y quizás hubiera producido algunos resultados interesantes a un nivel superior. Sea voluntario o no, creo que el hecho de que la mayoría de las medidas se hayan mantenido bastante intuitivas y comprensibles debe ser visto más como un resultado positivo que como algo negativo.

Además, en mi opinión, la metodología y las herramientas debían seguir las necesidades que se generaban a partir de los objetivos. Pero obviamente, estos objetivos sólo podían subsistir si existían algunas ideas o fundamentos para respaldar la investigación que se estaba iniciando. En mi caso, fueron las siguientes:

- La idea de re-conectar la noción intuitiva de información con el consumo de recursos o el esfuerzo computacional (o de razonamiento). En particular, la

ponderación del espacio y el tiempo tal y como expresa la función LT me parecieron una medida apropiada del esfuerzo.

- La idea de considerar la confirmación de la inferencia de una manera cuantitativa (pero no probabilística) bajo el siguiente supuesto: el valor de confirmación que una conexión deductiva proporciona debe ser mucho más grande que el que proporciona una conexión hipotética.
- La idea de que las hipótesis y teorías no deben ir acompañadas de un valor único de plausibilidad. Cualquier teoría de la confirmación útil debería dar un valor detallado para cualquier parte de la teoría. En particular, muchas de las diferencias entre teorías consilientes / intensionales y descripcionales / (parcialmente extensionales) se originan por el grado de uniformidad de la distribución de esta confirmación, una cuestión que ha sido ignorada en la literatura del aprendizaje automático.
- La idea de que la capacidad de razonamiento puede ser evaluada mediante medidas derivadas formal y computacionalmente. Más aún, la convicción de que no es necesario una referencia externa e inicialmente predeterminada para escalar la inteligencia de un agente, persona, animal o cualquier otro sistema cognitivo.

Veamos hasta qué punto la metodología e ideas anteriores han sido fructíferas:

4.1 Aportaciones Principales de la Tesis

Las contribuciones más importantes de esta tesis son:

1. Una medida de ganancia de información que ignora el tiempo $V(x|y)$, que representa la proporción de información de x que está *implícitamente* en y. ***Una nueva medida efectiva de ganancia de información computacional*** $G(x|y)$, que depende del esfuerzo computacional (tiempo y espacio) y se puede utilizar para medir la proporción de x que puede obtenerse fácilmente con la ayuda de y. En otras palabras, el grado de información de x que está explícitamente en y viene dado exactamente por la expresión $1 - G(x|y)$.

2. A partir de la ganancia de información computacional se definen ***Medidas de Ganancia de Representación y Optimalidad Representacional***. Se introducen una noción general de simplificación y un criterio de optimalidad representacional, así como medidas generales de optimización de sistemas y de poder sistemático.

3. ***Medida Uniforme para Inducción y Deducción***. En el caso de la inducción, la función G recoge perfectamente la idea Popperiana de informatividad. Asimismo, la deducción puede ser informativa y varias medidas se introducen para diversos paradigmas deductivos, especialmente teorías de primer orden. Se introduce una nueva noción de aprendizaje

auténtico, que asegura que se ha dado lugar un verdadero aprendizaje, independientemente de lo comprimible que sea la evidencia. Además, esta noción es aplicable a la deducción, mostrando que el aprendizaje no es sólo cosa de inferencia inductiva sino también de la deductiva. Se realiza una comparación con las ideas de Hintikka, estableciendo la correspondencia entre G y la información superficial y entre V y la información profunda.

4. ***Una nueva medida de refuerzo que cuantifica la propagación de confirmación en una teoría***. La teoría del refuerzo permite un tratamiento más detallado de las excepciones y proporciona diferentes valores para diferentes partes de una teoría, no la probabilidad única para toda la teoría que dan las distribuciones de probabilidad a priori.

5. ***La medida de refuerzo se comporta apropiadamente como medida de confirmación para diferentes procesos de inferencia*** como la inducción, la abducción, la analogía y la deducción, los cuales, a la larga, están envueltos en cualquier construcción de una teoría a partir de una evidencia. Nociones previamente vagas como la noción de 'consiliencia' de Whewell y la noción de inducción explicativa se formalizan fácilmente en este marco.

6. La Ganancia de Información y la medida de Refuerzo actúan como una pareja perfecta para discernir qué reglas deben dejarse explícitas en la representación de una teoría. A partir de aquí, ***se reconoce formalmente la necesidad de información intermedia y se deriva un criterio de olvido*** y se extiende para gestionar la evidencia pasada y ya explicada. En este contexto, la diferencia entre métodos anticipados (*eager*) y perezosos (*lazy*) se clarifica mediante la definición de un grado de pereza.

7. ***Se formaliza la idea de intensionalidad*** en términos de intolerancia o prohibición de excepciones, éstas vistas como partes extensionales o no validadas de una teoría. Es directamente aplicable a teorías lógicas y luego extendido a cualquier lenguaje descripcional, basada en una definición de subprograma. Esta idea se conecta también con la de 'consiliencia'.

8. Debido a los problemas del principio MDL, se introducen varios conceptos basados en la complejidad descripcional, tales como las descripciones proyectables y estables. La definición de una variante explicativa de la complejidad Kolmogorov permite la definición de ***una contrapartida explicativa al principio MDL.***

9. ***Un test no-antropomórfico de inteligencia***, basado en nociones computacionales y de teoría de la información, que puede significar un importante avance en la evaluación del progreso de la IA. Se introducen también varias particularizaciones para diferentes factores inductivos y deductivos. De este modo, la psicometría puede encontrar sus largamente

esperados fundamentos teóricos en la teoría de la información y la computación.

10. ***La aplicación de las medidas a diferentes tipos de sistemas lógicos y basados en el conocimiento***, tales como teorías Horn, bases de datos y sistemas software. Algunos aspectos del descubrimiento de conocimiento *nuevo* en bases de datos y la necesidad de información intermedia se entienden mejor en términos de ganancia de información. En el caso de sistemas software, la mantenibilidad se puede estudiar mediante la teoría del refuerzo, ya que cuanto más reforzado (usado) esté un componente software, menos probable será que se modifique en el futuro.

Aunque las medidas más importantes, ganancia de información computacional, refuerzo e intensionalidad, se definen independientemente, resultan ser útiles (solas o combinadas) para formalizar o comprender mejor conceptos muy diversos que han sido tradicionalmente bastante ambiguos: novedad, explicitez/implicitez, informatividad, sorpresa, interés, comprensibilidad, consiliencia, utilidad, incuestionabilidad, ...

Naturalmente, se analizan las relaciones entre estas medidas y otras más clásicas. La ganancia de información es análoga al *ratio* de ganancia de Quinlan para la inducción y similar a la información superficial y profunda de Hintikka para la deducción. La intensionalidad está estrechamente relacionada con la ganancia de información, ya que las descripciones extensionales no son nunca informativas. La noción de comprensión se relaciona también con la noción de incuestionabilidad, que se produce cuando no hay explicaciones alternativas. La medida de refuerzo y el principio MDL están también positivamente relacionados, aunque el refuerzo es más robusto a evidencias aleatorias, dando hipótesis más informativas. Algunos de estos resultados se han obtenido en general y otros se han particularizado para teorías lógicas o basadas en reglas.

Parte de estas contribuciones han aparecido en algunas revistas y conferencias, con el objetivo de dar difusión a este trabajo. En concreto, las contribuciones 1 y 3 aparecieron en Kurt Gödel Colloquium / Barcelona Logic Meeting (KGS'99), la contribución 4 aparecerá en International Journal of Intelligent Systems (IJIS), las contribuciones 7 y 9 fueron parcialmente incluidas (trabajo conjunto con N. Minaya) en International Symposium of Engineering of Intelligent Systems (EIS'98) y una versión de revista de ellas se ha enviado al número especial "Alan Turing and Artificial Intelligence" de Journal of Language, Logic and Information (JoLLI), la contribución 7 apareció en Model-Based Reasoning en Scientific Discovery (MBR'98) y será publicada en *Philosophica*, la contribución 8 se presentó en MBR'98 y está aceptada para *Foundations of Science* (en colaboración con I. García); finalmente, parte de las ideas de la contribución 10 sobre bases de datos se han utilizado en un artículo a ser presentado en el European Congress on Systems Science (ESS'99, trabajo conjunto con F. Alamagnac).

Como se verá en la siguiente sección, esta investigación ha abierto también muchas otras cuestiones que explorar, y espero que ellas muestren en mucha mayor amplitud la utilidad de las contribuciones anteriores.

4.2 Cuestiones Abiertas y Trabajo Futuro

Hay dos tipos de cuestiones suscitadas por esta tesis: de tipo técnico, que de algún modo muestran las limitaciones y problemas sin resolver dejados por esta tesis, y nuevas cuestiones de tipo teórico, que son, mayoritariamente, esperadas y deseables al final de cualquier tarea científica que también busca nuevos campos fértiles que explorar.

Entre las cuestiones técnicas abiertas, una fuente importante de desasosiego se encuentra en el coste computacional de las medidas que se han introducido. Como se dijo, la función de ganancia *G* es computable pero intratable. Se ha justificado, sin embargo, que no es extraño que sea así, porque, si fuera eficiente, podría ser usada para guiar procesos inductivos y deductivos, precisamente perturbando lo que se va a medir, el esfuerzo de un proceso de inferencia.

De todas maneras, se pueden obtener algunas aproximaciones inferiores de otras medidas que se han mostrado parcial y positivamente relacionadas con ellas, como la idea de intensionalidad. Decimos inferiores porque son útiles para descartar muchas hipótesis no informativas, pero no todas. Por ejemplo, el evitar extensionalidades (o explicitez) puede ser una buena heurística para obtener resultados informativos. De todos modos, la medida de refuerzo debe usarse siempre para evitar resultados informativos pero fantásticos.

Afortunadamente, la medida de refuerzo es computacionalmente factible, ya que el algoritmo presentado puede adaptarse para recalcular sólo las partes de la teoría que cambian después de que una revisión tiene lugar. Además, el criterio de olvido puede usarse para optimizar el uso de recursos espaciales, ya que la evidencia es generalmente tan grande que partes de ella deben olvidarse. La teoría del refuerzo permite saber cuáles deben olvidarse. No obstante, un estudio de otros algoritmos de propagación de validación que han sido introducidos en las últimas décadas en el área de redes neuronales artificiales (e.g. *backpropagation*) podría inspirar algunas mejoras o nuevos algoritmos para computar los valores de refuerzo.

Hay todavía, por supuesto, muchos viejos problemas que resolver (o aceptar), como la intratabilidad de la inducción (que a diferencia de la deducción, incluso se da para lenguajes muy restringidos). Esta intratabilidad (bien conocida desde los resultados de Gold) ha sido incluso respaldada por algunos resultados de esta tesis, ya que se ha demostrado que algoritmos de orden polinómico que trabajan a partir de los datos no pueden encontrar hipótesis informativas en general. Esto justifica el evitar métodos inductivos guiados por los datos y motiva el uso de técnicas más 'aleatorizadas', como los algoritmos genéticos, que pueden ser buenos inductores *en el*

caso medio. El uso de medidas de evaluación tales como el refuerzo pueden ser un criterio de optimalidad muy indicado para guiar el mecanismo de selección de este tipo de algoritmos. Esto está empezando a ser ensayado en (Hernández-Orallo and Ramírez-Quintana, 1998a, 1998b, 1999a). En el futuro, puede ser aplicado para la combinación de deducción e inducción para sistemas deductivos incompletos e ineficientes, especialmente presentes cuando aparece orden superior. Una estrategia de deducción puede beneficiarse en gran medida de técnicas inductivas.

Entre las cuestiones nuevas abiertas, subrayaría especialmente las implicaciones de la medición teóricamente fundada de las habilidades cognitivas que ha sido introducida en el capítulo 8. Muchas cuestiones fascinantes se abren a partir de la correlación de los clásicos tests psicométricos (IQ tests) y un C-test que está generado exclusivamente de las nociones de comprensibilidad e incuestionabilidad basadas en teoría de la información. Creo que la cuestión más trascendental que puede resolverse en el futuro es "*¿Cuán inteligentes somos?*" desde un punto de vista no-antropomórfico. Más aún, la psicometría puede comenzar un estudio teórico (ayudado por informática teórica) para estudiar formalmente (y no sólo experimentalmente) la independencia o dependencia de varios factores cognitivos. El hecho de que las habilidades deductivas correlen con las inductivas apoya la posición de que la inducción y la deducción no sean de ninguna manera procesos inversos.

A corto plazo, opino que la idea del test de Turing como un test práctico de inteligencia debería abandonarse, y sustituirse por tests computacionales y factoriales de diferentes habilidades cognitivas, una aproximación mucho más útil para el progreso de la inteligencia artificial.

Finalmente, un aspecto que sólo ha sido abordado parcialmente en esta tesis es la implicación de las medidas que se han introducido para la filosofía de la Ciencia. Aunque algunas nociones originalmente introducidas en este contexto han sido formalizadas, como la noción de consiliencia de Whewell, la noción de informatividad de Popper, la noción de descubrimiento y la distinción entre inducción explicativa y descriptiva, los resultados pueden ser explotados en mayor medida mediante una aproximación comprensiva de la informatividad y la confirmación en el contexto de la filosofía de la Ciencia. Lo mismo se aplica, en mayor o menor extensión, a la filosofía de las matemáticas, como se apunta en el capítulo 4.

En resumen, hay mucho trabajo teórico que hacer acerca de las cuestiones técnicas y teóricas anteriores. Por ejemplo, la conexión de la ganancia de información con la criptografía podría desvelar más detalles del comportamiento de *G*. Su relación con el aprendizaje PAC podría también sugerir la definición de una versión probabilística de la ganancia de información (aunque KT ya tiene en cuenta parcialmente esto porque pondera espacio y tiempo al considerar todas las posibles combinaciones). Otra línea de investigación teórica futura sería la relación y combinación de la teoría del refuerzo con la lógica difusa, o incluso con

aproximaciones neuro-fuzzy, que, en mi opinión, podría ser la propuesta más cercana a la manera en que la propagación del refuerzo se ha definido aquí.

Además, existe mucho trabajo experimental por hacer para explotar las posibles aplicaciones de las medidas que se han introducido en esta tesis. La reciente popularidad abrumadora del campo de los agentes racionales no ha sido acompañada por combinaciones exitosas y generales de diferentes procesos de inferencia, parcialmente debido a la falta de medidas que pudieran aplicarse consistentemente a todos ellos.

Las aplicaciones más inmediatas pueden aparecer a partir del uso de las medidas para la lógica de primer orden del capítulo 7, que pueden ser usadas directamente por la comunidad ILP. El capítulo 9 es también una sugerencia de aplicaciones potenciales para bases de datos, sistemas software y lenguaje natural, donde las ideas de informatividad e intensionalidad pueden ser cruciales para la comunicación y la comprensión.

A medida que más campos relacionados con los sistemas de información o de conocimiento, la inteligencia artificial, el procesamiento del lenguaje natural y el aprendizaje automático vayan integrando más técnicas de inferencia, las expectativas de aplicaciones de las medidas y otros conceptos que se han presentado en esta tesis serán todavía mayores.

4.3 Conclusiones Finales

A la vista de los resultados y las cuestiones abiertas que se han comentado en las secciones anteriores, creo que el objetivo principal de esta tesis: "*el estudio formal de la utilidad y resultado de la síntesis de conceptos en términos de ganancia de información y refuerzo en sistemas de inferencia, aplicables de manera consistente a la inferencia deductiva e inductiva*" ha sido alcanzado sustancialmente. El conjunto de medidas que han sido introducidas permiten un análisis detallado del valor de la salida de un proceso de inferencia con respecto a la entrada y el contexto (conocimiento previo o sistema axiomático), tanto en términos de informatividad como de confirmación.

En definitiva, el resultado más importante de esta tesis es el esclarecimiento de la relación entre las nociones de inferencia, información, comprensibilidad y confirmación. Como conclusión, la visión de inducción y deducción como procesos inversos en términos de ganancia de información se descarta definitivamente en sistemas no omniscientes y de recursos limitados, entre ellos los seres humanos y las computadoras.

5. Referencias

[Aha 1997] D.W. Aha, "Lazy Learning. Editorial" Special Issue about "Lazy Learning" AI Review, v.11, 1-5, Feb. (1997).

[Angluin 1988] Angluin, D., Queries and concept learning, Machine Learning 2, 4:319-342, 1988.

[Arcos and Plaza 1996] Arcos, J.L. and Plaza, E. "Reflection in Noos: An object-centered representation language for knowledge modelling" *Future Generation Computer Systems*, 1996.

[Armengol and Plaza 1994] Armengol, E.; Plaza, E. "Integrateing induction in a case-based reasoner" in J.P. Haton, M. Keane, and M. Manago (eds.) *Advances in Case-Based Reasoning*, number 984 in Lecture Notes in Artificial Intelligence, pp. 3-17, Springer-Verlag, 1994.

[Bar-Hillel and Carnap 1953] Bar-Hillel, Y.; Carnap, R. "Semantic Information" *British Journal for the Philosophy of Science* 4, 1953, 147-157.

[Bennett 1988] Bennett, C.H. "Logical depth and physical complexity", in Herken, R. "The universal Turing machine: a half-century survey" Oxford University Press, 1988, 227-258, 2nd Edition 1994.

[Bochenski 1965] Bochenski, J.M. "The methods of contemporary thought", Dordrecht, D. Reidel 1965, Spanish Translation, "Los métodos actuales del pensamiento", Rialp, Madrid, 1968.

[Bradford and Wollowski 1995] Bradford, P.G.; Wollowski, M., A Formalization of the Turing Test (The Turing Test as an Interactive Proof System), SIGART Bulletin, Vol. 6, No. 4, p. 10, 1995.

[Brand 1996] Brand, C. "The g Factor: General Intelligence and its implications" Wiley 1996

[Bratko and Dzeroski 1995] Bratko, I.; Dzeroski, S. "Engineering Applications of ILP" New Generation Computing, 13, 313-333, 1995.

[Brent 1993] Brent, J. "Charles Sanders Pierce: A life" Indianapolis: Indiana University Press 1993.

[Brockhausen and Morik 1997] Brockhausen, P. and Morik, K. "A multistrategy approach to relational knowledge discovery in databases. *Machine Learning* Journal, Vol. 27 (3), Kluwer, 1997

[Carnap 1950] Carnap, R. "Logical Foundations of Probability", Routledge & Kegan Paul, London, 1950.

[Carnap 1952] Carnap, R. "The Continuum of Inductive Methods" The University of Chicago Press, 1952.

[Chaitin 1982] Chaitin, G.J. "Gödel's Theorem and Information" Int. J. of Theoretical Physics, vol.21, no.12, pp. 941-954, 1982.

[Cohen and Nagel 1935] Cohen, M.R.; Nagel, E. "Introduction to logic and scientific method" in Spanish translation Cohen, M; Nagel, E. "Introducción a la lógica y al método científico" Amorrortu, Buenos Aires, 1968. Recent Re-edition: An Introduction to Logic. Indianapolis, Indiana: Hackett Publishing Company, 1993.

[Conklin and Witten 1994] Conklin, D.; Witten, I. H. "Complexity-Based Induction" *Machine Learning*, 16, 203-225, 1994.

[Cussens 1998] Cussens, James "Deduction, Induction and Probabilistic Support" Synthese Journal, 1998.

[Duc 1997] Duc, H.N. Reasoning about rational, but not logically omniscient, agents, Journal of Logic and Computation, Vol. 7, nº5, pp. 633-648, 1997.

[Dummett 1973] Dummett, M.A.E. "The Justification of Deduction" in Proceedings of the British Academy, LIX (1973), reprinted in *Truth and Other Enigmas*, London, Duckworth, pp. 290-318, 1978.

[Evans 1963] Evans, Thomas G. "A Heuristic Program to Solve Geometric Analogy Problems" in *Semantic Information Processing*, edited by Marvin Minsky, MIT Press, Cambridge, MA, 1968, Based on a PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1963.

[Flach 1995a] Flach, Peter "Conjectures. An inquiry concerning the logic of induction", Thesis Dissertation, Proefschrift Katholieke Universiteit Brabant, Tilburg, http://www.cs.bris.ac.uk/~flach/Conjectures/

[Girard et al. 1989] Girard, Jean-Yves; Taylor, Paul; Lafont, Yves "Proofs and Types" Cambridge Univ. Press, 1989.

[Gold 1967] Gold, E.M. "Language Identification in the Limit" *Inform. and Control*., 10, pp. 447-474, 1967.

[Good 1971] Good, I.J. "Twenty-seven principles of rationality" in *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott (eds.), Toronto: Holt, Rinehart and Winston 1971.

[Grégoire and Saïs 1996] Grégoire, E. and Saïs, L., Inductive reasoning is sometimes deductive, in M. Denecker, L. De Raedt, P. Flach and T. Kakas (eds) Working Notes of the ECAI'96 Workshop on *Abductive and Inductive Reasoning*, , 1996.

[Hempel 1943] Hempel, C.G. "A purely syntactical definition of confirmation" J. Symbolic Logic 6 (4): 122-143; 1943.

[Hempel 1945] Hempel, C.G. "Studies in the logic of confirmation" *Mind* 54 (213): 1-25; 54(214): 97-121, 1945.

[Hernández-Orallo 1998a] Hernández-Orallo, J. "Reinforcement Learning in Constructive Languages", Proceedings of CCIA'98 , pp. 264-272, Tarragona, 21 - 23 october de 1998

[Hernández-Orallo 1998b] Hernández-Orallo, J. "Formalising Consilience", in S.Rini, G.Poletti (eds.) Proceedings of the 1998 International Conference on Model Based Reasoning (MBR'98), pp. 25-27, Pavia 1998.

[Hernández-Orallo 1999a] Hernández-Orallo, J. "A Computational Definition of 'Consilience'", *Philosophica*, to appear.

[Hernandez-Orallo 1999b] Hernandez-Orallo, J. "Unified Information Gain Measures for Inference Processes", *5th Barcelona Logic Meeting and 6th Kurt Gödel Colloquium*, Barcelona, pp. 39-42, 1999.

[Hernández-Orallo 1999c] Hernandez-Orallo, J., "Universal and Cognitive Notions of 'Part'", to be presented at the *4th European Congress on Systems Science (ESS'99)*, 1999.

[Hernández-Orallo 1999d] Hernández-Orallo, J.: Constructive Reinforcement Learning, *Intl. J. of Intelligent Systems, to appear*. URL: http://www.dsic.upv.es/~jorallo/escritsa/IJISHern.ps.gz

[Hernández-Orallo 1999e] Hernández-Orallo, J, Unified Information Measures for Inference Processes, *Collegium Logicum* - Annals of the Kurt-Gödel-Society Vol. 4, Springer Verlag Wien, to appear.

[Hernández-Orallo 1999f] Hernández-Orallo, J. "Beyond the Turing Test", *Journal of Logic, Language and Information*, submitted.

[Hernández-Orallo and Alamagnac 1999] Hernández-Orallo, J.; Alamagnac, F. "Data Quality for Data-Mining" to be presented at the *4th European Congress on Systems Science (ESS'99)*, 1999.

[Hernández-Orallo and García Varea 1998b] J. Hernández-Orallo and I. García-Varea, "On Autistic Interpretations of Occam's Razor", in S.Rini, G.Poletti (eds.) Proceedings of the 1998 International Conference on Model Based Reasoning (MBR'98), pp. 25-27, Pavia 1998.

[Hernández-Orallo and García-Varea 1998a] J. Hernández-Orallo, J. and García-Varea, I. "Distinguishing Abduction and Induction under Intensional Complexity" in P. Flach and A. Kakas (eds.) *Proc. of the European Conference of Artificial Intelligence (ECAI'98) Ws. on Abduction and Induction in AI*, pp. 41-48, Brighton 1998.

[Hernández-Orallo and García-Varea 1999] Hernández-Orallo, J.; García-Varea, I. "Explanatory and Creative Alternatives to the MDL Principle", *Foundations of Science*, Kluwer, to appear,

[Hernández-Orallo and Hernández-Orallo 1993] Hernández-Orallo, J.; Hernández-Orallo, E.: *Programación en C++*, Paraninfo 1993, 2nd Ed., Intl. Thompson Publishers, 1995.

[Hernández-Orallo and Minaya-Collado 1998] Hernández-Orallo, J. "A Formal Definition of Intelligence based on an Intensional Variant of Algorithmic Complexity" Proceedings of Engineering of Intelligent Systems (EIS98), ICSC Academic Press 1998 .

[Hernández-Orallo and Pinto 1996a] Hernandez-Orallo, J.; Pinto, J. "Viabilidad de un Modelo de Conocimiento Falible en Cálculo de Situaciones" Departamento de Computación, Pontificia Universidad Católica de Chile, August 1996.

[Hernández-Orallo and Pinto 1996b] Hernandez-Orallo, J.; Pinto, J. "Especificación Formal de Protocolos Criptográficos en Cálculo de Situaciones" Departamento de Computación, Pontificia Universidad Católica de Chile, August 1996, to be published in Novatica, to appear 1999.

[Hernández-Orallo and Ramírez-Quintana 1998a] Hernández-Orallo, J. and Ramírez-Quintana, M.J. "Inductive Inference of Functional Logic Programs by Inverse Narrowing" J. Lloyd (ed) *Proc. JICSLP'98 CompulogNet Meeting on Comp. Logic & Machine Learning*, pp. 49-55, 1998.

[Hernández-Orallo and Ramírez-Quintana 1998b] Hernández-Orallo, J.; Ramírez-Quintana, M.J.: Inverse Narrowing for the Inductive Inference of Functional Logic Programs, in Freire-Nistal, J.L.; Falaschi, M.; Vilares-Ferro, M. (eds) Proc. 1998 Joint Conference of Declarative Programming., pp.379-392, 1998.

[Hernández-Orallo and Ramírez-Quintana 1999a] Hernández-Orallo, J. and Ramírez-Quintana, M.J. "A Strong Complete Schema for Inductive Functional Logic Programming", in Flach, P.; and Dzeroski, S. *Inductive Logic Programming'99* (ILP'99), in the Volume 1634 of the Lecture Notes in Artificial Intelligence (LNAI) series, Springer-Verlag 1999.

[Hernández-Orallo and Ramírez-Quintana 1999b] Hernández-Orallo, J. and Ramírez-Quintana, M.J. "Inductive Functional Logic Programming", 8th International Workshop on Functional and Logic Programming (WFLP'99), Grenoble, France, 28-30, June 1999.

[Hesse 1974] Hesse, M., *The Structure of Scientific Inference*, MacMillan, London, 1974.

[Hesse 1974] Hesse, M., *The Structure of Scientific Inference*, MacMillan, London, 1974.

[Hintikka 1970a] Hintikka, J. "On Semantic Information" in Hintikka, J.; Suppes, P. (eds.) D.Reidel Publishing Company, pp. 3-27, 1970.

[Hintikka 1973] Hintikka,J. "Logic, Language-Games and Information" The Calrendon Press, Oxford Univ. 1973.

[Holland et al. 1989] Holland, J.H.; Holyoak, K.J.; Nisbett, R.E.; Thagard, P.R. "Induction. Processes of Inference, Learning, and Discovery" The MIT Press 1989.

[Horvitz 1990] Horvitz, E. "Computation and Action Under Bounded Resources" PhD Dissertation, Stanford University, 1990.

[Johnson 1992] Johnson, W.L., Needed: A New Test of Intelligence, SIGART Bulletin, Vol. 3, No. 4, 7-9, October 1992, Editorial and Commentary.

[Kaelbling et al. 1996] Kaelbling, L.; Littman, M.; Moore, A.: Reinforcement Learning: A survey, *J. of AI Research*, 4, 237-285, (1996).

[Keynes 1921] Keynes, J.M., *A Treatise on Probability*, Macmillan, London 1921.

[Kirsh 1990] Kirsh, David "When Is Information Explicitly Represented?" in *Information, Language, and Cognition*, edited by Philip P. Hanson, Vancouver, University of British Columbia Press, 1990, 340-65.

[Kodratoff 1994] Kodratoff, Y. Guest Editor's Introduction "The Comprehensibility Manifesto" *AI Communications*, 7(2): 83-85, 1994.

[Kolmogorov 1965] Kolmogorov, A.N. "Three Approaches to the Quantitative Definition of Information" Problems Inform. Transmission, 1(1):1-7, 1965.

[Koppel 1987] Koppel, M., Complexity, Depth, and Sophistication, *Complex Systems* 1, 1087-1091, 1987.

[Koppel 1988] Koppel, M. "Structure", in Herken, R. "The universal Turing machine: a half-century survey" Oxford University Press, 1988, pp. 435-452, 2nd Edition 1994.

[Langley and Simon 1995] Langley, P., and Simon, H.A. "Applications of machine learning and rule induction" *Commun. ACM 38*, 11 (Nov. 1995), 55-64, 1995.

[Li and Vitányi 1997] Li, M.; Vitányi, P. "An Introduction to Kolmogorov Complexity and its Applications" 2nd Ed. Springer-Verlag 1997.

[Marcus et al. 1999] Marcus, G.F.; Vijayan, S.; Bandi Rao, S.; Vishton, P.M. "Rule Learning by Seven-Month-Old Infants" Science, January 1999, pp. 77-80

[Martín 1998] Martín, M. "Reinforcement learning for embedded agents facing complex tasks" Thesis Dissertation, Universitat Politècnica de Catalunya.

[Mitchell et al. 1991] Mitchell, T.M.; Allen, J, Chalasani, P.; Cheng, J. Etzioni, O.; Ringuette, M; Schlimmer, J. "THEO: A framework for self-improving systems" in Kurt VanLehn, editor, Architectures for Intelligence, pages 323-356, Lawrence Erlaum Associates, 1991.

[Moravec 1998] Moravec, H., *ROBOT: Mere Machine to Transcendent Mind*, Oxford Univ. Press, 1998.

[Muggleton 1994b] Muggleton, S. "Bayesian Inductive Logic Programming" in Cohen, W; Hirsh, H. (eds.) *Proceedings of the Eleventh International Machine Learning Conference* , San Mateo, CA, Morgan-Kaufmann, 371-379, 1994.

[Muggleton 1995a] Muggleton, S. "Inverse Entailment and Progol" New Generation Computing Journal 13:245:286, 1995.

[Muggleton 1995b] Muggleton, S. "Mode-Directed Inverse Resolution" in Kurukawa, K.; Michie, D.; Muggleton, S. (eds.) Machine Intelligence 14, Oxford University Press, 1995.

[Muggleton 1998] Muggleton, S., "Inductive logic programming: issues, results and the LLL challenge" in H. Prade, editor, Proceedings of ECAI98, page 697. John Wiley, 1998.

[Muggleton and Buntime 1988] Muggleton, S. and Buntime, W. "Machine invention of first-order predicates by inverting resolution" *Fifth International Conference on Machine Learning*. Morgan Kaufmann, 1988.

[Muggleton and De Raedt 1994] Muggleton, S. & De Raedt L. "Inductive Logic Programming — theory and methods" Journal of Logic Programming, 19-20:629-679, 1994.

[Muggleton and Page 1994] Muggleton, S.; Page, D. "A Learnability Model for Universal Representations" *Technical Report*, PRG-TR-3-94, Oxford University Computing Laboratory, Oxford 1994. URL: http://www.cs.york.ac.uk/~stephen/jnl.html

[Muggleton et al. 1992] Muggleton, S.; Srinivasan, A.; Bain, M. "Compression, significance and accuracy" in D. Sleeman and P. Edwards (eds.) *Machine Learning: Proceedings of the Ninth International Conference (ML92]*, pages 523-527, Wiley 1992.

[Mura 1990] Mura, A. "When Probabilistic Support is Inductive" Philosophy of Science 57, 278-289, 1990.

[Nake 1974] Nake F. "*Ästhetik als Informationsverarbeitung*" Springer 1974.

[Newell 1990] Newell, A. *Unified Theories of Cognition*, Cambridge, Mass.: Harvard University Press, 1990.

[Nilsson 1995] Nilsson, Nils J. "Eye on the Prize" AI Magazine, July 1995, also at http://robotics.stanford.edu/~nilsson/

[Plaza and Arcos 1993] Plaza, E. and Arcos, J.L. "Flexible Integration of Multiple Learning Methods into a Problem Solving Architecture" Report de Recerca IIIA 93/16 Octubre 1993, also appeared in Proceedings of the European Workshop on Knowledge Acquisition in 1994.

[Popper 1962] Popper, K.R. *Conjectures and Refutations: The Growth of Scientific Knowledge*, Basic Books, 1962.

[Popper 1963] Popper, K.R. *Conjectures and Refutations: The Growth of Scientific Knowledge*, Routledge and Kegan Paul, London, 1963.

[Popper and Miller 1983] Popper, K.R.; Miller, D. "A Proof of the Impossibility of Inductive Probability" Nature 302, 687-688.

[Popper and Miller 1987] Popper, K.R.; Miller, D. "Why Probabilistic Support is not Inductive" Philosophical Transactions of the Royal Society of London, A, 321, 569-591, 1987.

[Quinlan 1986] Quinlan, J.R. "Induction of Decision Trees" Machine Learning, 1:81-206, 1986.

[Quinlan 1990] Quinlan, J.R. "Learning Logical Definitions from Relations" *Machine Learning*, 5 (3): 239-266, 1990.

[Quinlan, 1993] Quinlan, J.R. "C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, C.A., 1993.

[Rissanen 1978] Rissanen, J.: Modelling by the shortest data description, *Automatica-J.IFAC*, 14, 465-471, 1978.

[Rissanen 1986] Rissanen, J., Stochastic complexity and modeling, Annals Statist. 14:1080-1100, 1986.

[Rissanen 1996] Rissanen, J.: Fisher Information and Stochastic Complexity, *IEEE Trans. Inf. Theory*, 1(42): 40-47, (1996).

[Russell and Wefald 1991] Russell, S.J. and Wefald, E.H. "Do the Right Thing: Studies in limited rationality" Cambridge, Massachusetts, MIT Press, 1991.

[Schmidhuber et al. 1997a] J. Schmidhuber, J. Zhao and M. Wiering, "Shifting Inductive Bias with Success-Story Algorithm, Adaptive Levin Search, and Incremental Self-Improvement" *Machine Learning*, 28, 105-132, (1997).

[Schmidhuber et al. 1997b] J. Schmidhuber, J. Zhao, N. Schraudolph. Reinforcement learning with self-modifying policies. In S. Thrun and L. Pratt, eds., Learning to learn, Kluwer, pages 293-309, 1997.

[Sebag and Rouveirol 1997] Sebag, M.; Rouveirol, C. "Tractable induction and classification in FOL" in Proceedings of IJCAI-97, 888-892, Morgan Kaufmann.

[Shanahan 1989] Shanahan, M. "Prediction is Deduction but Explanation is Abduction" in Proc. IJCAI'89, p.1055-1060, 1989.

[Simon 1982] Simon, H. "Models of Bounded Rationality" Cambridge, MIT Press 1982.

[Simon and Kotovsky 1963] Simon, H.; Kotovsky, K. "Human acquisition of concepts for sequential patterns" Psych. Review 70, 534-46, 1963.

[Solomonoff 1964] Solomonoff, R.J. "A formal theory of inductive inference" Inform. Contr. vol. 7, pp. 1-22, Mar. 1964; also, pp. 224-254, June 1964.

[Solomonoff 1978] Solomonoff, R.J. "Complexity-based induction systems: comparisons and convegence theorems" *IEEE Trans. Inform. Theory*, IT-24:422-432, 1978.

[Solomonoff 1986] Solomonoff, R.J. "The Application of Algorithmic Probability to Problems in AI" in L.N. Karnal; J.F. Lemmer(eds) *Uncertainty in AI*, Elsevier Science, pp.473-91, 1986.

[Sommer 1995a] Sommer, E.: "Fender: An approach to theory restructuring" in Proc of the European Conference on Machine Learning (ECML-95), 1995.

[Sommer 1995b] Sommer, E.: "An Approach to Quantifying the Quality of Induced Theories" in C. Nedellec (ed.), Proc. IJCAI'95 Workshop on Machine Learning and Comprehensibility, 1995.

[Suttner and Sutcliffe 1996] Suttner, C.B.; Sutchliffe, G. "The TPTP Problem Library", Tech. Univ. Munich, Germany, 1996

[Thagard 1978] Thagard, P. "The best explanation: Criteria for theory choice" Journal of Philosophy, 75, 76-92, 1978.

[Valiant 1984] Valiant, L. "A theory of the learnable". Communication of the ACM 27(11), 1134-1142, 1984.

[Whewell 1847] Whewell, W. "The philosophy of the inductive sciences" New York, Johnson Reprint Corp, 1847.

[Zilberstein 1995] Zilberstein, S. "Models of Bounded Rationality. A concept paper" AAAI Fall Symposium on Rational Agency, Cambridge, Massachusetss, November 1995.

[Zilberstein 1996] Zilberstein, S. "Resource-bounded reasoning in intelligent systems" Computing Survey 28(4), 1996.

[Zilberstein 1999] S. Zilberstein and the Resource-Bounded Reasoning Lab. "What is resource-bounded reasoning?" http://anytime.cs.umass.edu/Home.html